Verification of AI–based environmental forecasting systems: what can we do, what do we need to do, and what are the challenges?

Journal of the European Meteorological Society 4 (2026) 100032

Contents lists available at ScienceDirect

Journal of the European Meteorological Society

journal homepage: https://www.sciencedirect.com/journal/journal-of-the-european-meteorological-society

Verification of AI–based environmental forecasting systems: What can we do, what do we need to do, and what are the challenges?

Jochen Bröcker [a,b,c,*], Simon Driscoll [d], Tobias Necker [e,f], José Rodríguez [g], Helen Dacre [b], Natalie Harvey [b], Zied Ben Bouallègue [e]

[a] Department of Mathematics and Statistics, University of Reading, Reading, UK
[b] Department of Meteorology, University of Reading, Reading, UK
[c] Centre for the Mathematics of Planet Earth, University of Reading, Reading, UK
[d] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom
[e] European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK
[f] University of Vienna, Vienna, Austria
[g] Met Office, Exeter, UK

ARTICLE INFO

Keywords:
Artificial intelligence
Numerical weather forecasting
Forecast verification

ABSTRACT

Several institutions have released global medium–range meteorological forecasting models based on methods from machine learning, with training data provided by various reanalysis experiments. A proper and in-depth assessment of these models and the quality of their forecasts has yet to be carried out. Although in terms of simple and overall measures of skill such as mean square errors, AI-based forecasts clearly show very promising skill, we are just beginning to understand where and when these forecasts are useful and when they are not. Furthermore, while verification of meteorological forecasts has been subject to extensive (and still ongoing) research with a well established core methodology, it is not clear to what extent this methodology needs to be adapted or modified for AI–based models. Our paper aims to provide a vision on the verification of AI–based weather forecasts, identifying challenges, outlining important research questions, and laying the groundwork for a methodology to assess the quality of such forecasts.

1. Introduction

The start of the 2020s have seen considerable progress in the development of weather forecasting systems based on machine learning methodologies. From about 2022 onwards, several forecasting systems based on machine learning (ML) or artificial intelligence (AI) were presented which seem to achieve scores rivalling operational short and medium range weather forecasting systems such as the ECMWF high–resolution (deterministic) IFS forecasting system (see for instance Bi et al., 2023; Keisler, 2022; Lam et al., 2022; Pathak et al., 2022; Chen et al., 2023). At least, in relation to specific meteorological variables, there is evidence for AI–based forecasts exhibiting very competitive performance. Specifically, Keisler (2022) produces forecasts of specific humidity which appear to be more accurate than ECMWF's IFS system beyond day 3, using a graph neural network (GNN) architecture, which is also used in Lam et al. (2022) to produce forecasts which appear to outperform the IFS on several atmospheric variables. Pathak et al. (2022) combine a Fourier transform–based scheme with a vision transformer (ViT) to produce forecasts for 2 metre temperature which appear comparable in accuracy to the IFS. With similar transformer approaches, Bi et al. (2023) and Chen et al. (2023) produce forecasts for a range of variables that appear to be more accurate than IFS, and furthermore improve the scores compared to Lam et al. (2022) at longer lead times in particular.

In Ben Bouallègue et al. (2024), a systematic performance comparison of AI–based forecasts in an operational–like context is presented. The authors focus on the PanguWeather ML model in Bi et al. (2023), which is freely available for non–commercial use. Both PanguWeather and the operational NWP forecast are initialised on the same initial conditions, unlike in previous studies where the AI–based forecasts were typically initialized on ERA5 atmospheric conditions. The authors of Ben Bouallègue et al. (2024) apply both standard verification techniques that are normally used at ECMWF, and innovative statistical tools suggesting new avenues for the verification of AI–based forecasts. Several important advances in this domain have emerged since then, such as the advent of ensemble–based models (for example GenCast,

* Corresponding author.
E-mail addresses: broecker@reading.ac.uk (J. Bröcker), sd2136@cam.ac.uk (S. Driscoll), tobias.necker@ecmwf.int (T. Necker), jose.rodriguez@metoffice.gov.uk (J. Rodríguez), h.f.dacre@reading.ac.uk (H. Dacre), n.j.harvey@reading.ac.uk (N. Harvey), zied.benbouallegue@ecmwf.int (Z. Ben Bouallègue).

https://doi.org/10.1016/j.jemets.2026.100032
Received 23 June 2025; Received in revised form 29 January 2026; Accepted 9 February 2026
Available online 6 March 2026

a probabilistic diffusion ensemble AI–based model introduced in Price et al., 2025), expansion of datasets for potential benchmarking, and growing use of precipitation as a diagnostic target (see also Section 5). These trends underscore a shift toward more comprehensive, diverse, and observationally grounded verification practices that extend beyond conventional metrics and are increasingly aligned with both regulatory and scientific expectations.

Moreover, Bonavita (2024) pointed to the limitations of AI–based models in representing physical dynamical balances, arguing that current models achieving good error scores despite lacking physically realistic behaviour highlights a need for verification beyond root mean square error (RMSE) and toward more physically and structurally meaningful metrics. Liu et al. (2024) propose advancing beyond traditional point-to-point comparisons and towards adaptive spatial, temporal, and intensity-aware corrections. Whilst focused on nowcasting, their proposal demonstrates the potential for more flexible and physically grounded metrics in the space of AI–based forecasts for better capturing the characteristics of the system. They offer one of many potential directions that could be encompassed within broader verification frameworks as those set out here.

In this rapidly evolving context of AI–based forecasts, this present paper aims to articulate a vision for the field of verifying AI–basedweather forecasts. We discuss foundational ideas for methodologies and frameworks for the verification of AI–based forecasts, and identify key research questions in this area.

This paper is strongly informed by discussions held at the first Ver–AI workshop, which took place at the University of Reading on 24th and 25th of June 2024 and brought together international experts across the areas of operational weather forecasting, forecast verification, and AI–based forecasts. The full workshop programme can be found in Appendix A. Key discussions around the following themes then influenced the present paper:

Theme 1: statistical evaluation of AI–based forecasts, including ensembles. What are the statistical properties of AI–based forecasts in terms of reliability, resolution, correlations, signal-to-noise ratios, and events of extreme magnitude and duration? How do we assess these? How would the methodology differ from classical weather forecasts?

Theme 2: ML benchmarks for weather and climate problems. Benchmarks allow AI researchers without domain expertise to make significant contributions. What are good design principles for such ML benchmarks?

Theme 3: physical properties and interpretability of AI–based forecasts. To what extent do AI–based forecasts exhibit the typical atmospheric balances such as mass conservation or geostrophic balance? How realistic is the representation of complex spatio–temporal meteorological patterns such as storms, droughts or blocking events in AI–based forecasts?

. The present work expands these themes into a vision paper for verification of AI based weather forecasts. Sections 2–4 discuss points that emerge from these three themes, respectively, partially with conclusions. Section 4 furthermore lists additional considerations for verifying AI–based forecasts beyond these themes. Section 5 concludes with a few take–home points aimed at practitioners, a very brief overview over the most recent developments that have taken place in the last year or so in this rapidly evolving field, and finally with a variety of open research questions.

2. Long term statistical properties of AI–based forecasts, including ensembles

The long-term statistical properties of AI–based forecasts are not only to be understood as performance fact sheets of a given forecasting system. More generally, the statistical properties bear on the explainability of AI–based forecasts, and the metrics and verification methods used in the assessment need to be selected with that in mind. It is also clear that a lot can be learned and incorporated from existing NWP assessment practices and wisdom (see for instance Jolliffe and Stephenson, 2012; Wilks, 2006, Ch. 7). Furthermore, there are a number of reasons as to why the evaluation of AI–based forecasts require additional ideas and methods—these will be discussed below.

An overarching issue which touches simultaneously upon several of the points discussed below is statistical significance. Evaluations so far have been over periods of one or a few years at best. Given the relevant timescales of atmospheric processes however, these evaluations, although useful, have to be treated with care, and conclusions based on these evaluations must be seen with caution. (This point is not specific to AI–based forecasts of course, and similar caveats would apply to classical forecasting systems.) Statistical techniques to assess significance over longer periods, such as bootstrapping or cross validation over multiple years, will therefore play a very important role in the assessment of AI–based forecasts, at least for some time.

Do AI–based forecasts require a different verification methodology?. Despite the wide range of new and additional possibilities for verifying AI–based forecasts, there is the danger that a new methodology becomes driven by the needs and constraints of AI–based forecasting systems, rather than driving the development of AI–based forecasts. The methodology needs to be interpretable in meteorological terms and thus equip AI–based forecasts with accountability. The enormous amount of existing knowledge, understanding, and theory about meteorology, weather prediction, and geophysical model design needs to be harnessed for further development of AI–based forecasts. This requires substantial knowledge transfer and communication between different communities. Sharing expertise on how the AI–based forecasts are developed and operate would help researchers in the NWP community to understand where their mathematical and physical knowledge is most usefully applied to develop AI–based forecasts further.

Events of extreme duration and magnitude. Weather extremes are an important area for verification of AI–based forecasts. As in machine learning the data is split into training and test (and potentially validation) subsets, therefore special attention is required to ensure that extremes are adequately represented in those data sets. Certain extreme events however, although rare locally, occur with larger frequency globally, such as tropical cyclones. Such events could be artificially introduced into datasets that have too few of them. However, lessons learned from Large Language Models suggest that training on synthetic data can also introduce biases or degrade performance elsewhere (see e.g. Shumailov et al., 2024).

Forecast verification therefore may have to account for the use of synthetic data, given that the data no longer reflects the statistical properties of the real weather. A potentially related problem with extremes in AI–based models however seems to be the overly smooth fields produced by these forecasts (in particular those using transformer architectures, see Bonavita, 2024), and addressing this issue will be important in the future (see also Section 4).

This should include guidance and methods to treat or exclude the smoothing effect in verification.

Large ensembles. The benefit of AI–based models is that, once trained, they are fast and cheap to run, and taking advantage of large ensembles is likely to be beneficial in the representation of extreme events and their evaluation. Establishing criteria and methods for probabilistic verification of large ensembles, and in particular of extreme events in large ensembles would be valuable (Bröcker, 2018; Bröcker and Ben Bouallègue, 2020; Necker et al., 2024) This should also include quantifying the benefit of increased ensemble sizes which requires a good understanding of the ensemble size dependence of applied verification methods. Another interesting opportunity created by our ability to produce very large numbers of ensemble members is to have ensemble systems

where the number of members is not fixed but depends on the dynamical situation. The ensemble sizes could thus become "adaptive" depending on the predictability of the state of the atmosphere at the time. To make use of this new possibility, a proper understanding and evaluation methodology of flow–dependent forecast skill in AI–based models is needed.

Precipitation. Precipitation forecasting is relatively new in AI–based forecasts. A more complete evaluation of such important meteorological variables is urgently needed as we require a better understanding of these in particular in the context of AI–based forecasts. Although traditional evaluation methods could be applied to both classical as well as AI–based models alike, issues like data sparsity in regions without observations may pose challenges in the training of AI–based models. This could be particularly important in the evaluation of recently developed models that include precipitation (see Section 5).

New assessment metrics. Statistical assessment of AI–based forecast properties should involve an evaluation of the extent to which they inherit statistical characteristics from training data, which may vary depending on initialisation methods and data sources such as reanalysis. Again as current splits allow for only short verification periods, one may consider methodologies for extending verification periods, possibly through retrospective analyses or synthetic data generation. Indeed, this could be applied to the above for extremes, but serves as a wider point too.

Verification in latent space. A new approach to evaluating AI–based models which is not available in classical NWP models is the possibility of evaluation in latent space. AI–based models transform the data from the physical space onto a latent space, which typically has a lot fewer dimensions. The possibilities and benefits of doing the verification in that space should be explored. To assess the potential of latent space methods for verification, the model forecast would not be mapped into physical space but left in latent space. Meanwhile the current observation (which, for all currently available AI–based models, is an analysis from a classical NWP model) would be mapped into latent space using the map that is provided by the AI–based model. The comparison with the model forecast would then take place in that space. This approach to evaluation does not rely on the fact that "observations" are currently analyses from a classical NWP model and could be performed even if AI–based models use actual observations.

Latent space analysis might have drawbacks, however, for instance in models such as Pangu–Weather which have a high dimensional latent space with configurable dimensionality. In such cases, latent space analysis may be less straightforward. It is also possible that one might need to perform latent space analysis across layers or blocks of layers which also can increase the dimensionality once more. It may still nonetheless remain promising even in the these complex settings, as it could offer an alternative view on forecast similarity and structure bypassing challenges associated with physical output fields.

As an aside, Neighborhood Verification Methods (Ebert, 2009) are an example of a classical NWP verification method that would be very interesting to explore on the context of AI–based forecasting.

Instationarities and climate change. For prospective AI–based forecasts covering longer leadtime such as annual or even decadal, instationarities related to climate change become important (Rackow et al., 2024). Currently, CO_2 forcing is not typically included in the training data but it would clearly be important for such models.

3. ML benchmarks for weather and climate problems

Benchmarks are a very important paradigm in AI and Machine Learning. They allow AI researchers without domain expertise to make significant contributions by developing AI–models and architectures that perform well at least against the benchmark. A recent example is WeatherBench 2, see Rasp et al. (2024). In order for the achieved performance to generalise, the benchmarks need to satisfy a number of requirements, and it is important to consider what is actually needed from benchmarking. Identifying "good" or desirable design principles for such ML benchmarks is therefore crucial to further developing AI–based forecasts (Dueben et al., 2022).

Better benchmark datasets. Coordination across various international stakeholders is required to create new benchmark datasets (or improve existing ones) for various parameters (of various weather and climate parameters but also of parameters related to impacts (for example precipitation, renewables, extremes, climate parameters, and damage due to high winds, flooding, and storm surges). To ensure that contributing to benchmark datasets is easy and straight forward, the community should aim for open–source verification tools and platforms that facilitate community contributions.

In addition to suitable datasets, a benchmark dataset should provide a suite of experimental designs for testing the models. (Thus the name benchmark *dataset* is not fully adequate.) The experiments should include (but not be limited to) scenarios of particular end user interests, such as European windstorms, tropical cyclones, or heat waves. The ECMWF severe event catalogue (Magnusson, 2019) could be a source of inspiration that should be supplemented with suitable experimental design.

Verification of models against actual observations (rather than only analysis fields) should also be included, although it is understood that this is not a trivial thing to include as it requires additional domain expertiseadditional domain expertise on the part of the user is typically required to properly interprete the observational data as well as the results from any verification analysis using such data.

Fairness of evaluation and standardisation of methodology and software. In order for fairness across models a framework for comparison is needed, and the characteristics a fair evaluation framework has to exhibit need to be identified.the evaluation methodology to be fair across models, a framework for comparison is needed, and the properties which such a fair evaluation framework ought to have need to be identified. Fair comparisons should allow for a comparative evaluation of models which essentially predict the same environmental variables (e.g. the global atmosphere) but which use slightly different configurations, for instance different grids. In the case of AI–based models, different training data sets might similarly impede a direct comparison of models. Fair scores for instance (introduced in Ferro, 2014) have been developed to account for differences in ensemble sizes when comparing classical physics–based models; they can obviously be applied to AI–based models. But more generally, fair scores provide an example of how to suitably generalise a concept so as to use it for comparing the performance of models with different configurations in a fair manner. Other examples need to be developed, such as a suite of standardised plots and code which could then be referenced as an "official" set of tests that has been designed and assessed for fairness, and be included in the benchmarking.

Explainable AI and assessing explainable behaviour. Accompanying the use of benchmarks, with standard scenarios to assess the capability of models, a core part of any assessment of the AI–based forecasts should be using Explainable AI (or XAI) techniques. Explainability tools should be integrated into the verification and benchmarking pipeline to compare equally and understand model behaviour. Explainability in AI–based systems will be required by law (European Union, 2024) as was noted during the Ver–AI workshop itself by Anna-Louise Ellis (Met Office).[1]

[1] The EU AI act recital 27 states that "Transparency means that AI–systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an

Standardisation of tests with explainable behaviour of AI–based models would help with those models comply with 'regulations': The accessibility of forecast data shapes model adoption, leading to questions about the future role of national meteorological services and the influence of big tech companies entering the field. Ethical considerations, such as those outlined in the Hiroshima Process International Code of Conduct for Advanced AI Systems, (Group of Seven , 2023b), (see also the related G7 Leaders' Statement, Group of Seven, 2023a), emphasise the need for transparent and explainable AI–based models, although methods for doing this are not well developed and AI–based models may be held to a higher standard than traditional NWP models which may be problematic.

Diversity of data sets. There is lack of diversity in both the training and evaluating datasets, which mainly consists of ECMWF Reanalysis v5 (ERA5) data. A more diverse portfolio of reanalysis data sets should be used in training and evaluation. (Examples of potentially suitable but apparently underused data sources are the NOAA 20th Century Reanalyses or NASA's Integrated Multi-satellitE Retrievals for GPM (IMERG) product.) Yet as there is no obvious economic incentive, private companies are unlikely to do this. As there is potentially great scientific insight to gain, this is something that academic institutions should pursue, in collaboration with public operational weather centres.

Evaluation on different datasets however might lead to unfair comparisons. "Standardising" verification data can help but also risks promoting overfitting to benchmarks. It is not clear however how to standardize verification datasets so that model comparisons remain fair, without encouraging models to overfit to benchmark data.

Related to this, the consistency of the global observations used was discussed intensively. Currently, AI–based models are trained predominantly on ERA5 reanalysis data, and therefore a AI–multi–model ensemble might not have the same diversity as witnessed in climate models. Yet as more AI–based models arise, and different data sources become available for training, this might change, leading to greater diversity in AI–based models and to potentially greater benefit of using AI–multi–model ensembles. AI–based models trained solely on observational data rather than reanalyses (see for instance McNally et al., 2024; Alexe et al., 2024; Allen et al., 2025) are of course particularly interesting in this regardwould of course further increase this diversity and thus be particularly interesting in this regard. Again, in order to assess the diversity and realise the potential benefits, a framework for comparison against standardised benchmarks is required.

Authorities, institutions, and private vs public research. In principle, the scientific community is incentivised to create benchmarks as there is scientific value in them. It is not clear though how to practically facilitate not only the design of benchmarks but also associated scholarly research into the design of appropriate benchmarks. It is not clear, nor was there consensus in the Ver–AI workshop, as to whether this would require a top–down approach, involving funding agencies and strategic research plans of bigger institutions, or whether we can rely on the initiative of a few people to initiate this, in the hope that it catches on and evolves over time with the input and contribution of the community.

Even though there do not seem to be any obvious economic incentives to creating (publicly available) benchmarks, as discussed above in relation to datasets, one may argue that benchmarking initiatives should be driven by the scientific community, rather than by private enterprises. The Coupled Model Intercomparison Project (CMIP) groups provide good examples of how this could work as they are very active communities.

However, since the private sector is currently one of the largest drivers of AI–based forecasting, it is unclear if such an approach would be also adopted by the private sector as well. Bauer (2024) for instance emphasises the need for strong public–private collaboration. Such collaborations have achieved, for example, the development of a successful prescriptive TOGAF standard for enterprise architecture (Josey, 2018). How to strike a good balance between the potentially conflicting interests in such collaboration is noted as an open question for future discussion at the end of our paper.

An important activity of the scientific community would be to organise periodical conferences or workshops to maintain the initiative, especially when the benchmarking is expected to change in a field that is evolving very rapidlybenchmarking is expected to change quickly in a field that is evolving very rapidly overall.

End users and the economy. In addition to scientific institutions and private companies, end users are obvious stakeholders that needs to be incentivised to contribute to benchmarks. For this to work it is clearly necessary that the benchmarks are useful tools for the contributors. This bears on the employed metrics for instance; presumably metrics of interest to typical users will relate to societal impacts and vulnerabilities, like precipitation, wind and solar radiation for renewable energy industries, or to extreme events like droughts, storm surges, etc.

Regulatory concerns. The proliferation of AI–based models underscores the need for regulatory frameworks to distinguish between reliable and unreliable models—this was echoed by participants of the Ver–AI workshop. Domain–specific expertise is important here, with considerations of brand identity influencing model choice of end users. WeatherBench 2 represents a community–driven effort where experimental AI–based models have submitted forecasts for evaluation. Yet challenges persist as models are evaluated in an automated fashion solely through metrics without human intervention. Addressing this gap could involve creating consumer–like reports to independently assess model suitability, potentially involving professional societies to validate both public and private sector AI–based models for weather and climate forecasts.

4. Physical properties and interpretability of AI–basedforecasts

A major feature of classical physics–based atmospheric and ocean models is their interpretability in physical terms. Although this may sound almost like a tautology, the physical interpretability (such as the well–known balances, e.g. hydrostatic or geostrophic) is the reason for much of the confidence we have in the adequacy of those models for the purpose of simulating weather and climate. Despite the extensive assessment of forecast performance which is carried out on a continuous basis, model development and maintenance is still very much informed by physical insight.

An important example of physical interpretability are the typical atmospheric balances such as mass conservation and geostrophic balance, among others, that physics–based atmospheric and ocean models exhibit. Whether and to what extent AI–based forecasts also reproduce those balances is currently subject to intensive research (Bonavita, 2024). It is not clear how realistic and robust complex spatio–temporal meteorological patterns such as storms, droughts or blocking events are represented in AI–based forecasts, and whether there is a sense in which the key features of such events are somehow encoded in the model.

XAI again. Interpretability is crucial for understanding how AI–based forecasting models learn and make predictions, and techniques in explainable AI can uncover the "reasoning" behind AI–based predictions. For example, measures of attention can be used to reveal which input features or regions of the atmosphere the AI–based model focuses on when making predictions. Techniques like deep latent space analysis can determine how AI–based models organise and utilise information, which aids in diagnosing biases and can potentially be used to identify previously unknown relationships in the atmosphere.

AI system, as well as duly informing deployers of the capabilities and limitations of that AI system"

An important question in the context of explainable AI is how to probe whether an AI–based model is getting the right answer for the right reason. While traditional physics–based models represent explicit physical equations which adhere to the typical atmospheric balances (such as mass conservation and geostrophic balance), AI–based models learn from data without enforcing such constraints directly. This can lead to issues where AI–based models are not guaranteed to conserve energy or maintain other physical balances, which then makes it difficult to identify the reasons behind potential forecast errors or failures (or indeed successes).

Stable evaluation benchmarks in the face of rapid development. Evaluation of AI–based models is a moving target as development of these models is very quick so the community needs to have tools to able to assess these things that can be applied to newly published models or model versions in both a timely and systematic manner. A set of standard scores for instance that can be tracked over time would provide a very useful picture of the AI–based model development. Likewise, basic physical consistency checks can be used to validate and gain trust in AI–based model output, but appropriate tools need to be provided to calculate these.

Physical conservation, physical structure, physical realism. Classical NWP models are designed to possess many of the fundamental balances and conservation laws of the climate system as innate behaviour, while for AI–based forecasts these are acquired traits. Failure of a classical NWP model to exhibit a balance it has been designed to obey is a sure sign that the model does not work as intended, and that an error has been made at some point in the design or implementation.If a classical NWP model fails to exhibit a physical balance that was originally built into the design of that model, we can be certain that the model does not work as intended, and that an error has been made at some point in the design or implementation. In AI–based forecasts the situation is more complicated. Assessing whether a model shows certain balances (or more generally exhibits certain dynamical behaviour) must be done through evaluation, including but not limited to case studies. Yet in AI–based forecasts there is generally less reason to believe that such a case study generalises than in classical models based on physical reasoning. We can check whether the model shows the required behaviour in the available case studies but unlike as in classical models there are no structural guarantees, and we cannot be sure an AI–based model will behave correctly in all future cases.

In Charlton-Perez et al. (2024), AI–based models and NWP models are compared in their ability to forecast the case study of Storm Ciarán, a rapidly deepening extratropical cyclone in November 2023 that brought record low pressure and severe winds the UK and Europe. Whilst the AI–based models generally got the cyclone's track correctly, they were unable to capture the intensity of wind speeds in the storm. Furthermore, whilst the AI–based models did well on the large scale structure of the storm, the physical structure and realism of high impact mesoscale features was insufficient, both when compared to observations as well as to NWP models, even those with similar resolution. This left a mixed message for the performance of AI–based models. Charlton-Perez et al. (2024) noted that storms with genesis similar to Ciarán's are common even in a relatively short ERA5 record. It remains an open question as regards how AI–based models perform for storms that are have less usual dynamics.It remains an open question as to how AI–based models perform for storms that exhibit very unusual dynamics and are therefore rare in the training data record.

A "critical theory" of AI–based model verification. Any purveyor of forecast information (private or public) clearly has an incentive to overrate the performance of the forecasts, especially if the purveyor is unlikely to face repercussions if (typically after some time) the forecast turns out to perform worse in the face of real data than was previously advertised. Using "objective" metrics or scores does not completely remove the ability to produce overly optimistic performance assessment. The assessment may focus on performance measures, forecast variables, and forecast periods which present the forecasting system at its best and then unduly generalise the results (or fail to highlight the limitations of the assessment). Furthermore, also "objective" metrics may favour certain properties of the forecast which are not necessarily of interest or of benefit to the user. Although these concerns apply to all kinds of forecasts, the fact that AI–based forecasts tend to exhibit a smoothing effect which gets worse with lead time is probably a manifestation of this.

It is clear that AI–based models should also be assessed with criteria against which they do not perform very well. It is not so clear however how such criteria would be identified. Geographical areas for instance that have a problem (e.g. rain over desert areas that would not be expected) are typically identified through visual inspection by skilled individuals. Therefore it does not seem straight forward to put this into a general framework or set of metrics.

Community building. There is a need for revising the agenda of the forecast verification research community, for widening the community across several discipline boundaries, and for potentially adopting new paradigms and methods. Learning from other domains and in particular ML experts clearly needs to be facilitated. Furthermore, forecast verification is important in other fields (such as finance and econometrics) and we might need to learn from those communities, in particular as AI–based forecasts are becoming more prevalent in other fields, too.

Verification of AI–based forecasts is now part of the mission of the WMO Working Group on Forecast Verification Research for instance, which may help further developments within the community. As already mentioned in Section 3, operational weather centres should continue to support the community and research into the verification of AI–based forecasts, if only to ensure a level playing field with their private and public counterparts, the list of which is likely to grow in the future.

Potential activities could include:

1. Workshops and annual meetings;
2. Regular publications of benchmarks and recommendations for verification;
3. Cross–disciplinary workshops to explore transferable verification techniques, and to connect with ML experts to foster collaboration and knowledge exchange;
4. Organize hackathons, prediction competitions, and collaborative projects to encourage community participation;
5. Attract joint funding and secure resources in particular to coordinate and develop benchmark datasets.

5. Take–home points and current developments

It is hard to draw final conclusions in the wider context of AI–based weather models due to the rapid evolution of the field. We have however identified key take–home points we feel will be integral to verification of AI–based forecasts in the near future:

1. AI–based forecasts require at least as comprehensive evaluation as classical physics–based forecasts; arguably, the evaluation has to be more comprehensive as our understanding of AI–based forecasts is currently much less complete than for classical forecasts.
2. Physical realism, balances, and dynamical consistency is not an innate property of AI–based forecasts but at best an acquired trait. Statistical evaluation and case studies can assess the dynamical properties of AI–based forecasts but as long as we do not fully understand the internal mechanisms of AI–based forecasts, we have less guarantee that a specific AI–based model is able to reproduce the dynamics not only of common and frequent features but also in less common situations. Advances in explainable AI in AI–based forecasting systems would help building this confidence.
3. Benchmarks have proved to be invaluable for progress in various machine learning contexts, and weather forecasting is no exception.

There is much room for further development though, and an enormous opportunity for the community of academics and practitioners. A diverse set of easily available verification benchmarks is bound to spurn further experimentation and development in AI–based forecasts. Finally, AI will increasingly be subject to legislatory regulation, and the community has an important role to play with regards to both coping with and shaping that legislation.

Current developments, open questions and future research. This paper discusses needs for verification of AI–based forecasts, building on discussions about the verification of AI–based models which took place in June 2024, in the context of a fast-evolving field of research. Since then, progress has been made in AI for weather forecasting, with several noteworthy developments. For example, AI–based models for ensemble forecasting have been developed that generate ensemble members with very little smoothing effect (Lang et al., 2024b; Price et al., 2025). Ensembles with large sizes have been explored in Mahesh et al. (2025), and precipitation is becoming a diagnostic variable more commonly issued by AI–based models (Lam et al., 2023; Lang et al., 2024a). Additional benchmark datasets have been developed with various targets and observation reference (Jin et al., 2024; Radford et al., 2025), and AI–based models trained directly from observations have been developed (McNally et al., 2024; Alexe et al., 2024; Allen et al., 2025). Along with these developments goes tremendous progress specifically in the field of evaluation of AI–based forecasts. Future Ver–AI workshop editions will also serve as a platform to continue the advancement of verification of AI–basedforecasts, as outlined in this vision paper.

We leave the reader with some open questions (sometimes with additional notes) to consider for future research that may play central roles in verification of AI-forecasting.

1. Is there a circularity in using one and the same scoring rule to first train the AI–based forecast model and then evaluate it? Would using a variety of metrics be preferable, in particular *different* metrics for training and verification, or would this constitute an unfair disadvantage?
2. AI is increasingly being used in longer-term simulations such as climate emulators or hybrid AI-physics models on longer horizons. Here they may be increasingly used in policy decisions where verification is also harder. What adaptations are needed to verify data-driven models and hybrids when used in longer-term seasonal or climate simulations? What do we expect from AI–based forecasts on longer time scales such as seasonal or climate?
3. In the days when computing power was much more expensive, multi–model ensembles were sometimes the only way of generating heterogenous forecasts (hence the epithet “poor man’s ensemble”). The ability to produce very large ensembles very cheaply is one of the main advantages typically advertised of AI–based forecasts (and explored e.g. in Mahesh et al., 2025), and thus the need for poor man’s ensembles seems to have disappeared (see Bröcker and Kantz, 2011, for the implications of exchangeable vs non–exchangeable ensemble members). There might still be an advantage though of using multi–AI–based forecast model ensembles, for instance to asses structural uncertainties. The argument is that the individual strengths of each model will be present in the “supermodel”, although this clearly needs to be investigated further.
4. Selz and Craig (2023) found that AI–based forecasts do not exhibit rapid error growth, and therefore do not simulate sufficiently the chaotic dynamics of the atmosphere. Understanding better how errors propagate in AI–based models and how to verify the sensitivity to initial conditions of these models could be a goal of future research.
5. Training on synthetic data may help improve performance of AI–based forecast models, for instance with regards to extreme events. Such gains might come at a price, so how should verification standards account for the use of synthetic data, given its potential to improve extreme-event prediction but also risk degrading performance elsewhere?

CRediT authorship contribution statement

Jochen Bröcker: Writing – review & editing, Writing – original draft, Investigation, Conceptualization; **Simon Driscoll:** Writing – review & editing, Writing – original draft, Investigation; **Tobias Necker:** Writing – review & editing, Writing – original draft, Investigation; **José Rodríguez:** Writing – review & editing, Writing – original draft, Investigation; **Helen Dacre:** Writing – review & editing, Writing – original draft, Investigation; **Natalie Harvey:** Writing – review & editing, Writing – original draft, Investigation; **Zied Ben Bouallègue:** Writing – review & editing, Writing – original draft, Investigation.

Data availability

No data was used for the research described in the article.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the speakers of the Ver–AI 2024 workshop (see Appendix A) for their contributions and fruitful discussions. SD acknowledges support of the project SASIP funded by Schmidt Sciences (Grant number G-24-66154) for his time at the University of Reading, and from Schmidt Sciences, LLC, for his time at the Institute of Computing for Climate Change (ICCS) within DAMTP, Cambridge. NJH is funded by the Natural Environment research council (NERC) grant number NE/X018555/1. J.M. Rodríguez was funded by the Met-Office Climate Science for Service Partnership (CSSP) China project under the International Science Partnerships Fund (ISPF). The workshop was organised by the Centre for the Mathematics of Planet Earth (CMPE) at the University of Reading; CMPE acknowledges financial support from the University of Reading and the ClimTip project (Horizon Europe UKRI underwrite Grant agreement 101137601).

We thank the reviewers and the editor for their invaluable comments, points for future discussion and enriching our paper overall.

Appendix A. Programme of the Ver–AI workshop 2024

Time	Speaker	Title
June 23rd:		
13.10	Simon Lang (ECMWF)	The AIFS: ECMWF's ML forecasting system
13.50	Simon Driscoll (Reading)	Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán
14.20	(Discussion)	
15.40	Tobias Necker (ECMWF)	The fractions skill score for ensemble forecast verification
16.20	Anna-Louise Ellis (Met Office)	Ethics, "Explainability" and XAI
16.50	Lewis Blunn (Met Office)	The use of citizen weather stations and urban flux observations in training and evaluating machine learning models
June 24th:		
9.30	Zied Ben Bouallegue (ECMWF)	Forecast realism: a new verification mantra?
10.00	Martin Leutbecher (ECMWF)	Ensemble size dependence of the logarithmic score for forecasts issued as multivariate normal distributions
10.40	(Discussion)	
13.00	Nkuiate Harris Sop (Exeter)	Evaluating Probabilistic Forecasts in the Presence of Observation Error
13.40	Etienne Roesch (Reading)	Computational reproducibility
15.30	Cedric Mesnage (Exeter)	Stability of AI models and transfer learning
Poster	Jose M. Rodriguez (Met Office)	Development of systematic errors in the East Asian summer monsoon
Poster	Yoshinori Tashiro (Reading)	Evaluation of AI-driven weather forecasts of extreme wind events in Europe

References

Alexe, M., Boucher, E., Lean, P., Pinnington, E., Laloyaux, P., McNally, A., Lang, S., Chantry, M., Burrows, C., Chrust, M., Pinault, F., Villeneuve, E., Bormann, N., Healy, S., 2024. GraphDOP: towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations. https://arxiv.org/abs/2412.15687.

Allen, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W.P., Andersson, T.R., Herzog, M., Lane, N.D., Chantry, M., Hosking, J.S., Turner, R.E., 2025. End-to-end data-driven weather prediction. Nature 641, 1172–1179. https://doi.org/10.1038/s41586-025-08897-0

Bauer, P., 2024. What if? numerical weather prediction at the crossroads. J. Eur. Meteorol. Soc. 1, 100002. https://doi.org/10.1016/j.jemets.2024.100002

Ben Bouallègue, Z., Clare, M. C.A., Magnusson, L., Gascon, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J.S., Lang, S. T.K., et al., 2024. The rise of data-driven weather forecasting: a first statistical assessment of machine learning–based weather forecasts in an operational-like context. Bull. Am. Meteorol. Soc. 105 (6), E864–E883.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q., 2023. Accurate medium-range global weather forecasting with 3d neural networks. Nature 619 (7970), 533–538.

Bonavita, M., 2024. On some limitations of current machine learning weather prediction models. Geophys. Res. Lett. 51 (12), e2023GL107377. e2023GL107377 2023GL107377. https://doi.org/10.1029/2023GL107377

Bröcker, J., 2018. Assessing the reliability of ensemble forecasting systems under serial dependence. Q. J. R. Meteorolog. Soc. (accepted). https://doi.org/10.1002/qj.3379

Bröcker, J., Ben Bouallègue, Z., 2020. Stratified rank histograms for ensemble forecast verification under serial dependence. Q. J. R. Meteorolog. Soc. 146 (729), 1976–1990. https://doi.org/10.1002/qj.3778

Bröcker, J., Kantz, H., 2011. The concept of exchangeability in ensemble forecasting. Nonlinear Process. Geophys. 18 (1), 1–5. https://doi.org/10.5194/npg-18-1-2011

Charlton-Perez, A.J., Dacre, H.F., Driscoll, S., Gray, S.L., Harvey, B., Harvey, N.J., Hunt, K. M.R., Lee, R.W., Swaminathan, R., Vandaele, R., Volonté, A., 2024. Do AI models produce better weather forecasts than physics-based models? a quantitative evaluation case study of storm ciarán. npj Clim. Atmos. Sci. 7, 93. https://doi.org/10.1038/s41612-024-00638-w

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., Quyang, W., 2023. Fengwu: pushing the skillful global medium-range weather forecast beyond 10 days lead. arXiv preprint arXiv:2304.02948.

Dueben, P.D., Schultz, M.G., Chantry, M., Gagne, D.J., Hall, D.M., McGovern, A., 2022. Challenges and benchmark datasets for machine learning in the atmospheric sciences: definition, status, and outlook. Artif. Intell. Earth Syst. 1 (3), e210002. https://doi.org/10.1175/AIES-D-21-0002.1

Ebert, E.E., 2009. Neighborhood verification: a strategy for rewarding close forecasts. Weather Forecast. 24 (6), 1498–1510. https://doi.org/10.1175/2009WAF2222251.1

European Union, 2024. European union artificial intelligence act. Accessed May 22nd, 2025. https://artificialintelligenceact.eu/recital/27.

Ferro, C. A.T., 2014. Fair scores for ensemble forecasts. Q. J. R. Meteorolog. Soc. 140 (683), 1917–1923. https://doi.org/10.1002/qj.2270

Group of Seven, 2023a. G7 leaders' statement on the hiroshima AI–process. Accessed May 22nd, 2025. https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process.

Group of Seven, 2023b. Hiroshima process international code of conduct for advanced AI–systems. Accessed May 22nd, 2025. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems.

Jin, W., Weyn, J., Zhao, P., Xiang, S., Bian, J., Fang, Z., Dong, H., Sun, H., Thambiratnam, K., Zhang, Q., 2024. Weatherreal: a benchmark based on in-situ observations for evaluating weather models. https://arxiv.org/abs/2409.09371.

Jolliffe, I.T., Stephenson, D.B. (Eds.), 2012. Forecast Verification; A Practicioner's Guide in Athmospheric Science. John Wiley & Sons, Ltd., Chichester. 2nd ed..

Josey, A., 2018. An Introduction to the TOGAF® Standard, Version 9.2. White Paper W182. The Open Group. Published 16 April 2018. https://publications.opengroup.org/w182.

Keisler, R., 2022. Forecasting global weather with graph neural networks. arXiv preprint arXiv:2202.07575.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., Battaglia, P., 2023. Learning skillful medium-range global weather forecasting. Science 382 (6677), 1416–1421. https://doi.org/10.1126/science.adi2336

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al., 2022. Graphcast: learning skillful medium-range global weather forecasting. arXiv preprint arXiv:2212.12794.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C.A., Lessig, C., Maier-Gerber, M., Magnusson, L., Ben Bouallègue, Z., Prieto Nemesio, A., Dueben, P.D., Brown, A., Pappenberger, F., Rabier, F., 2024a. AIFS – ECMWF's data-driven forecasting system. https://arxiv.org/abs/2406.01465.

Lang, S., Alexe, M., Clare, M. C.A., Roberts, C., Adewoyin, R., Ben Bouallègue, Z., Chantry, M., Dramsch, J., Dueben, P.D., Hahner, S., Maciel, P., Prieto Nemesio, A., O'Brien, C., Pinault, F., Polster, J., Raoult, B., Tietsche, S., Leutbecher, M., 2024b. AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. https://arxiv.org/abs/2412.15832.

Liu, J., Zhang, T., Chen, Y., Wang, R., Wang, M., Wang, S., Xu, T., Zhao, C., Chen, X., 2024. A new verification approach for nowcasting based on intensity and spatial-temporal feature correction. Sci. Rep. 14 (1), 30531.

Magnusson, L., 2019. ECMWF severe event catalogue for evaluation of multi-scale prediction of extreme weather. ECMWF Tech. Memoranda 851. 110.21957/i2pbf6pe

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Harrington, P., Kashinath, K., Kurth, T., North, J., OBrien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., Willard, J., 2025. Huge ensembles part II: properties of a huge ensemble of hindcasts generated with spherical fourier neural operators. https://arxiv.org/abs/2408.01581.

McNally, A., Lessig, C., Lean, P., Boucher, E., Alexe, M., Pinnington, E., Chantry, M., Lang, S., Burrows, C., Chrust, M., Pinault, F., Villeneuve, E., Bormann, N., Healy, S., 2024. Data driven weather forecasts trained and initialised directly from observations. https://arxiv.org/abs/2407.15586.

Necker, T., Wolfgruber, L., Kugler, L., Weissmann, M., Dorninger, M., Serafin, S., 2024. The fractions skill score for ensemble forecast verification. Q. J. R. Meteorolog. Soc. 150 (764), 4457–4477. https://doi.org/10.1002/qj.4824

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al., 2022. Fourcastnet: a global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al., 2025. Probabilistic weather forecasting with machine learning. Nature 637, 84–90. https://doi.org/10.1038/s41586-024-08252-9

Rackow, T., Koldunov, N., Lessig, C., Sandu, I., Alexe, M., Chantry, M., Clare, M., Dramsch, J., Pappenberger, F., Pedruzo-Bagazgoitia, X., Tietsche, S., Jung, T., 2024. Robustness of AI-based weather forecasts in a changing climate. https://arxiv.org/abs/2409.18529.

Radford, J.T., Ebert-Uphoff, I., Stewart, J.Q., Musgrave, K.D., DeMaria, R., Tourville, N., Hilburn, K., 2025. Accelerating community-wide evaluation of AI models for global weather prediction by facilitating access to model output. Bull. Am. Meteorol. Soc. 106 (1), E68 – E76. https://doi.org/10.1175/BAMS-D-24-0057.1

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., et al., 2024. Weatherbench 2: a benchmark for the next generation of data-driven global weather models. J. Adv. Model. Earth Syst. 16 (6), e2023MS004019.

Selz, T., Craig, G.C., 2023. Can artificial intelligence-based weather prediction models simulate the butterfly effect? Geophys. Res. Lett. 50 (20), e2023GL105747. e2023GL105747 2023GL105747. https://doi.org/10.1029/2023GL105747

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y., 2024. Ai models collapse when trained on recursively generated data. Nature 631 (8022), 755–759.

Wilks, D.S., 2006. Statistical Methods in the Atmospheric Sciences. Vol. 59 of International Geophysics Series. Academic Press, Oxford. 2nd ed.