Adv. Sci. Res., 23, 9–19, 2026
https://doi.org/10.5194/asr-23-9-2026

Advances in Science & Research
Open Access Proceedings

EMS Annual Meeting: European Conference for Applied Meteorology and Climatology 2025

Forecaster perspectives on model quality and verification at the Norwegian Meteorological Institute

Teresa Remes[1], Morten Køltzow[1], Andrew Singleton[1], Lene Østvand[1], Mette Sundvor Skjerdal[2], Bjørn Gilje Lillegraven[1], and Gunnar Noer[3]

[1]Development Centre for Weather Forecasting, Norwegian Meteorological Institute, Oslo, Norway
[2]Development Centre for Weather Forecasting, Norwegian Meteorological Institute, Bergen, Norway
[3]Development Centre for Weather Forecasting, Norwegian Meteorological Institute, Tromsø, Norway

Correspondence: Teresa Remes (teresa.remes@met.no)

Received: 30 March 2026 – Revised: 17 June 2026 – Accepted: 25 June 2026 – Published: 5 August 2026

Abstract. Successful application of numerical weather prediction (NWP) and emerging data-driven forecast models depends not only on their objective performance but also on how forecasters interpret, trust and communicate their outputs and uncertainties. This study presents insights from a series of verification workshops conducted with operational duty forecasters at the Norwegian Meteorological Institute. The workshops aimed to foster dialogue between model developers and forecasters, improve verification systems, and create evaluation procedures that serve both model development and operational forecasting. The workshops explored how forecasters evaluate model quality, identified parameters and products that are challenging to forecast or evaluate, and explored necessary tools for building trust in model products, including data-driven forecasts. Forecasters emphasized the importance of personal operational experience, peer interaction, and case-based verification in building knowledge about model quality. The findings indicate that forecasters' experience-based verification complements quantitative objective verification and plays an important role in forecasters' decision-making. Forecasters expressed a need for verification systems tailored to operational situations, such as specific events or similar past cases, and highlighted the importance of practical tools that can be used during shifts. They also stressed the need for reliable verification that includes comparisons with traditional NWP models for building trust in data-driven models, particularly for extreme events. Overall, the study highlights the importance of the communication of model strengths and weaknesses and the co-development of tools to support operational forecasting. It provides a basis for further development of verification practices and enhancing collaboration between forecasters and model development.

1 Introduction

Weather forecasting is based on constantly evolving numerical weather prediction (NWP) models with different spatial and temporal resolutions and time scales. More recently, data-driven models, based on machine learning, have been developed with a view towards becoming used in operational settings. Successful application of these models in forecasting depends not only on their objective, quantitative skill, but also on how forecasters understand, trust and communicate the model outputs and their uncertainties. This means that model development must be done together with evaluation and verification approaches that reflect real operational needs.

Skilful forecasts can be considered to have three types of goodness, as defined by Murphy (1993): consistency, quality and value. Consistency refers to agreement between the forecaster's judgement and the forecast issued, quality refers to the correspondence between forecasts and observations, and value reflects the benefit of the forecast to the user in their decision-making context. Importantly, a forecast can be consistent and of high quality but still provide little value if it does not influence decisions or actions. In this paper, we use the term model quality mainly to refer to how well model forecasts correspond to observations, which is close to Murphy's definition of forecast quality. Because the study focuses on forecaster perspectives, we also examine forecasters' perceptions of model quality and how these perceptions are formed in practice.

Forecast value is particularly important when forecasts are used to support public safety. Despite the increasing automation and availability of advanced forecast products based on NWP and data-driven models, the role of the forecaster remains critical in operational meteorology. Forecasters assess and interpret available model data and observations; apply heuristic reasoning and intuition, especially when guidance is uncertain (Doswell, 2004); and decide when and how to issue warnings that can help to protect lives and property. Forecasters add value by integrating model information with real-time observations, by applying their experience and situational awareness, by identifying critical issues, and by having an ability to make well informed decisions (Doswell, 2004; Kain et al., 2003; Steenburgh et al., 2013).

User perspectives and the value of weather forecasts for decision-making have caught considerable attention over recent years. Often the focus has been on end users and downstream applications, such as forecasts for the general public (Padilla et al., 2021), road work (Fundel et al., 2019), air quality (Ainslie et al., 2022), and farmers (Vigna et al., 2024). These studies highlight how probabilistic or impact-based forecasts are interpreted and used in specific decision making contexts, as well as how user-perspectives can be used to draw recommendations for implementation and improvement of forecast products.

In contrast, less attention has been paid to weather forecasters' perspectives despite their central role at national weather services issuing real-time forecasts and weather warnings. These forecasters are often the direct point of contact for most weather forecast users, in terms of both automatised model products and individualised and dedicated adjusted products. Forecasters' subjective confidence in a forecast influences how uncertainty is communicated to users, and even qualitative expressions of confidence can support better decision-making and trust (Padilla et al., 2021). In addition, the value of forecast products depends on how forecasts are interpreted and applied, even among professional users (Ainslie et al., 2022). Very few studies have specifically investigated forecasters' model use and needs (e.g., Kain et al., 2003; Demuth et al., 2020) despite the growing recognition that their perspectives are essential for improving model evaluation, communication and decision support.

At the Norwegian Meteorological Institute (MET Norway), operational NWP systems are monitored through several verification activities. These include quarterly summary reports of model performance at Norwegian weather stations (e.g., Lillegraven et al., 2025a, b, c, 2026), internal monitoring of key parameters against reference models, and evaluation reports written by forecasters of issued extreme warnings. In addition, model developers run dedicated verification scripts, often focusing on specific processes (e.g, sea ice, land surface and cloud processes), and forecasters can provide feedback directly to model developers during shifts. Taken together, these verification and monitoring activities provide substantial information about model performance. However, they have largely been developed for model monitoring and reporting purposes, and not systematically designed to support forecasters in their operational decision-making. This distinction is important because verification needs differ between users: model developers often need diagnostics of systematic errors and model changes, whereas forecasters need information that supports real-time assessment of forecast trustworthiness and uncertainty (Pagano et al., 2024). More broadly, developing useful verification capability in national meteorological and hydrological services requires attention not only to metrics but also practices that support communication, continuous improvement and user needs (Pagano et al., 2025).

In addition to the differences in verification needs, summary scores and traditional point-to-point verification methods have known limitations in diagnosing model behavior (Casati et al., 2008). In fact, summary measures cannot possibly inform on the whole complex forecast–observation relationship. Moreover, traditional verification is typically based on comparing gridbox values at the nearest model grid point to observations from in-situ weather stations. For variables with high spatial variability, such comparisons can be misleading. Small spatial displacements can lead to double penalties, where high-resolution models are penalized for placing a feature close to, but not exactly at, the observed location. Spatial verification approaches have been developed to address such limitations (Gilleland et al., 2009; Dorninger et al., 2018). In addition, point observations may not be representative of the model grid box or of the spatial scale relevant for the forecast (Göber et al., 2008; Køltzow et al., 2019; Remes et al., 2025). This includes, for example, representativeness errors related to strong small-scale variability of wind speed in complex terrain (Køltzow et al., 2019, 2022) and the undercatch of precipitation in gauge observations (Køltzow et al., 2020; Bouallegue et al., 2020). However, spatial verification approaches and other methodological improvements in verification, such as accounting for observation representativeness or implementation of model comparisons against remote sensing data, are not easily implemented in the current operational system. As a result, several parallel solutions have emerged to meet specific verification needs, but these do not necessarily serve forecasting very well.

Objective verification, nevertheless is generally seen as essential for judging when and where model output can be trusted, beyond subjective experience alone (Demuth et al., 2020). For verification to support operational forecasting, it should be aligned, not only with model development needs, but also with forecasters' needs. This underlines the importance of understanding how forecasters perceive model quality and how verification information could more effectively support their work.

In this study, we present insights from a series of verification workshops involving operational duty forecasters (i.e., meteorologists on shift responsible for, amongst other duties, issuing weather warnings) at MET Norway. The workshops were designed to build dialogue between model developers and forecasters in order to develop verification systems and create evaluation procedures that benefit both model development and forecasting. Specifically, the primary goal of these initial workshops was to better understand how forecasters evaluate model quality, which parameters and products are seen as the most challenging, and what conditions and tools are necessary for building trust in model products, including new data-driven forecasts.

2 Methods

The study was conducted at MET Norway, which operates three main forecasting offices in Oslo, Bergen and Tromsø, all of which undertake core forecasting tasks, like issuing weather warnings, in addition to aviation forecasting. Additionally, each of the forecasting offices have distinct (regional) responsibilities: forecasters in Tromsø (northern Norway) are specialized in Arctic forecasting, those in Bergen (west coast of Norway) in maritime forecasting, and the forecasters responsible for public service weather forecasts on national television are based exclusively in Oslo (southeastern Norway). The development group is mainly located in Oslo but also includes personnel who divide their time between forecasting at one of the main forecasting office and development work. To foster collaboration and dialogue between the development and forecasting departments regarding model quality and verification, a series of workshops was organised at these forecasting offices.

The workshops were held on 3–4 April 2024 in Tromsø, 3–4 December 2024 in Bergen and 10 and 12 June 2025 in Oslo. At each location, the workshops were conducted on two separate days with identical content to better fit the participants' shift schedule. All forecasters in these locations were invited to participate, and a total of 50 forecasters attended. Participants represented a range of experience levels, expertise and forecasting responsibilities, including safety, aviation, offshore/energy, and media.

Each workshop lasted approximately 3 h. They began with presentations on verification methods and a review and discussion of the most recent season's verification results for the specific region where the workshop was held. This was followed by a structured discussion. Data collection included facilitator notes from group discussions and individual feedback written on sticky notes. The qualitative data were later grouped into broad categories to find common themes. Data from all offices were combined to gain an overall understanding of the forecaster perspectives. This also limits the potential selection bias due to different forecasting responsibilities across regions.

A set of predefined question guided the structured discussion:

1. How/where do you gain knowledge about how good the models are?
2. How well are different parameters/phenomena forecasted?
3. Are there parameters, forecasts, or products at MET Norway that should be evaluated, evaluated more, or evaluated differently than today?
4. What is required for you to trust AI-generated forecasts (in the same way as traditional NWP forecasts)? (asked only in Bergen and Oslo)
5. What knowledge or tools related to forecast quality can help you during your shift?
6. How can knowledge about forecast quality be communicated (internally and externally)?

To capture individual perspectives, participants were invited to provide anonymous written feedback on the questions on sticky notes. For Question 2, participants were specifically asked to write a parameter, phenomenon, and/or model on a sticky note and place it on a board. The board was divided into two axes: forecasters' perceived quality of the available NWP guidance for selected parameter or phenomenon (x axis), and their confidence in their own judgement of that perceived quality (y axis). The x axis was intended to capture whether forecasters considered the available model guidance to perform poorly or well in an operational sense, based on their experience and expectations, rather than relative to a fixed reference. Afterwards the sticky notes were photographed, a grid was added over the board, and each comment was assigned coordinates based on its placement. This process is depicted in Fig. 1, which shows illustrations of (a) the board after participants had placed their sticky notes, and (b) the grid used to categorise the responses. The coordinates were used as an approximate categorisation of perceived model quality and confidence, not as a quantitative verification measure.

Operational forecasters at MET Norway have access to three main atmospheric NWP systems. For the Scandinavian domain (mainland Norway, Sweden and Finland), the Meteorological Co-operation on Operational NWP Ensemble Prediction System (MEPS; Eresmaa et al., 2026) based on the HARMONIE-AROME model system, is used. MEPS has a horizontal resolution of 2.5 km and consists of 30 ensemble members. MET Norway runs MEPS in collaboration with Swedish, Finnish and Estonian national meteorological institutes, but the scope of this study is limited to Norway. In addition, MET Norway operates AROME-Arctic (Müller

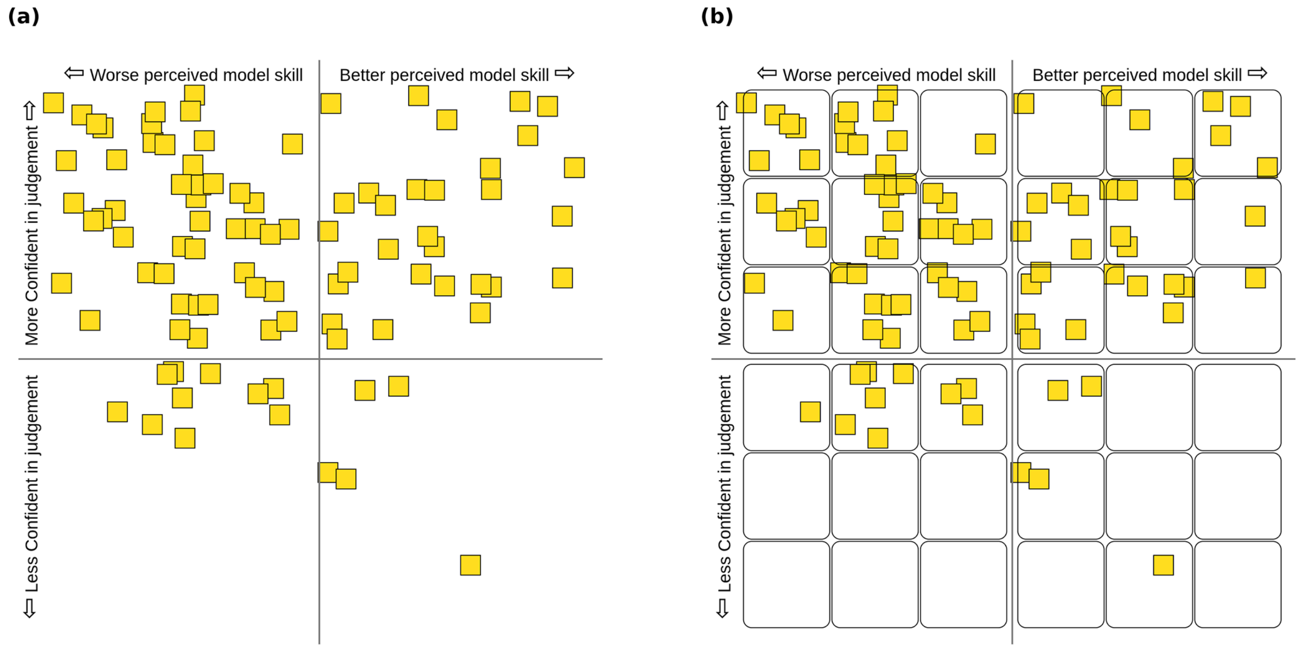

Figure 1. Illustration of data collection method for Question 2 ("How well are different parameters/phenomena forecasted?"): **(a)** Example of sticky notes placed on the board by participants at one forecasting office, with forecasters' perceived quality of available NWP guidance on the x axis and confidence in their own judgement on the y axis; and **(b)** board with an overlaid grid used to quantify the responses from that forecasting office. The placement reflects forecasters' subjective assessment in an operational context and should not be interpreted as an objective verification score.

et al., 2017) also based on the HARMONIE-AROME model system, with a horizontal resolution of 2.5 km and a domain centered over the Barents Sea and includes northern mainland Norway and the Svalbard archipelago. Forecasters also use products from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS), including the high-resolution deterministic forecast (HRES) and the ensemble prediction system (ENS; 51 members, approximately 9km resolution; ECMWF, 2024), which are typically used for medium- to long-range forecasts. During the workshops, participants were free to select which model systems or products they wished to discuss.

3 Results

The results are presented in six thematic areas corresponding to the predefined questions discussed during the workshops.

3.1 Forecasters' sources of knowledge

Forecasters rely on four primary sources to build their knowledge about model quality. First, personal experience gained during shifts emerged as the most important source of information. By repeatedly comparing observations with model products in real weather situations, forecasters build an understanding of model performance. Second, discussion with peers is seen as essential. These conversations often help forecasters to learn from others' experiences and refine their own judgements, especially regarding past weather events. Third, different organised interactions such as formal training, workshops and colloquia – often focused on quantitative verification – are mentioned as a source of information. Lastly, forecasters independently use a range of online verification tools, verification reports and research articles to learn about model quality.

3.2 Perceptions of model quality

The number of sticky notes placed on the board for question 2: "How well are different parameters/phenomena forecasted?" (Fig. 1a) was aggregated depending on which grid square (Fig. 1b) they were within, and the totals across all workshops are shown in Fig. 2. The results should be interpreted as a summary of forecasters' self-assessed perceptions of model quality and confidence in their own judgement, rather than as an objective verification of model performance. Forecasters expressed strong confidence in their ability to assess model quality, with 81 % of the written responses being above the x axis. At the same time, 60 % of the responses indicated a negative perception of model quality, mostly with high confidence in this judgement, whereas 40 % of the responses indicated positive perception of model quality but generally with lower confidence. Forecasters were free to select which models they wished to comment on. Only 44 % of

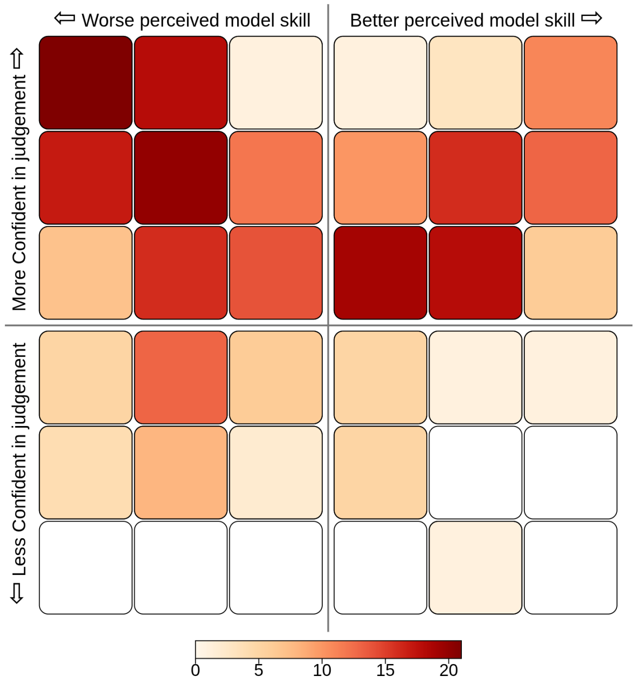

Figure 2. Distribution of comments from all locations based on perceived quality (x axis) and confidence in judgement (y axis).

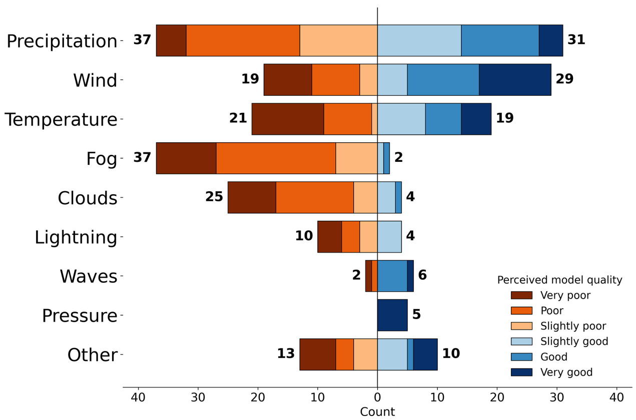

Figure 3. Distribution of comments indicating poor and good model quality by parameter. The "other" category includes rare mentions of phenomena such as turbulence, icing, atmospheric stability, and the jet stream. Colors show the level of perceived model quality. Parameters are ordered by the total number of comments.

the written responses referred to a specific NWP model, 58 % of these commented on MEPS, 31 % ECMWF IFS, and 7 % AROME-Arctic, while the rest concerned wave models (not the focus in this study).

Most of the responses regarded precipitation, wind, temperature, fog and clouds, with fewer mentions of lightning, waves, and atmospheric pressure (Fig. 3). Individual comments on other phenomena such as turbulence, icing, atmospheric stability and jet stream were mentioned less frequently. Precipitation, which gathered the most comments of all variables, received mixed views. Large-scale, stratiform precipitation was considered well forecasted, particularly over land, while showers and intense precipitation events were seen as less well captured. Wind forecasts were generally perceived as well predicted, particularly for moderate wind speeds and over oceans, but lower quality was perceived in complex terrain, at individual stations, and during strong wind events. This perception should be interpreted in light of the more limited observational data available over the ocean compared with land stations. At MET Norway, forecasts over the ocean can be assessed using offshore platform observations (Køltzow et al., 2024) and scatterometer winds (Køltzow et al., 2019), while buoy observations are sparse. In addition, many of the maritime forecasters are in direct contact with (specialized) offshore users of their product and receive feedback. The perception of higher quality over oceans may also reflect the less disturbed wind field over open water compared with complex terrain. Temperature forecasts were also seen as mostly reliable, with the exception of cold winter minima, which were described as systematically forecasted with a warm bias. Fog and clouds were frequently identified as poorly forecasted and this was associated with a relatively high confidence in judgement. Lightning also received more negative than positive responses. In contrast, waves and pressure were generally perceived well forecasted.

It is important to note that these perceptions are influenced by the weather regimes forecasters commonly encounter, which naturally shape their evaluations of model performance. At the same time, many of the comments provided on the sticky notes included descriptions of specific weather situations, demonstrating forecasters' awareness of how model performance varies between regimes.

Overall, these results highlight that forecasters have developed confident perceptions of model quality across different parameters and weather phenomena. At the same time, uncertainty was also expressed for some specific phenomena that are often difficult to verify systematically, such as fog, precipitation phase and lightning. Several of the perceived model strengths and weaknesses correspond to known model limitations documented in verification reports and previous studies (e.g., Frogner et al., 2019; Køltzow et al., 2019).

3.3 Evaluation needs

Forecasters were asked which parameters, forecasts, or products at MET Norway should be evaluated, evaluated more, or evaluated differently compared with today. Their responses highlighted a wide range of topics, with a strong emphasis on warnings of extreme or high impact weather for spatial areas, aviation-related products, convective precipitation, and fog and clouds. For warnings (yellow, orange, or red depending on severity and expected impact) issued by forecasters, it was noted that verification is currently done mostly for the orange and red warnings. Yellow warnings are issued for expected minor consequences, such as local power outages, delays in traffic or strong winds that might make mountain activities dangerous. Orange warnings indicate substantial impacts for many people with potential risk to life and property, and red warnings are issued when expected consequences are extreme with likely widespread damage and risk to life. In Norway, the specific criteria like thresholds for submitting warnings are adapted to different regions, based on expected damage. Forecasters expressed a desire for systematic evaluation across all warning levels, including yellow warnings.

They also suggested verifying aspects such as lead time, spatial extent, thresholds, and how well the warnings matched observed impacts, taking into account population density, infrastructure and vulnerability. Good archiving and database tools were mentioned as essential for tracking and comparing past events.

For aviation, forecasters emphasised the need for more detailed verification of aviation-specific model output parameters such as visibility, cloud base height, icing, and turbulence. They also highlighted the need for including the vertical structure of the atmosphere, such as temperature and humidity profiles, in the verification and stressed that verification should reflect the conditions most relevant for operational decision-making.

Convective precipitation was another recurring theme, and participants noted that it is less reliably forecast than stratiform rain. Forecasters suggested that verification should distinguish these weather regimes to capture differences in model performance. Similarly, fog, visibility, and cloud cover were commonly identified as particularly difficult to predict and also to verify. These parameters are critical in many operational contexts, especially in coastal regions, and forecasters requested more site-specific evaluations and targeted metrics to address these challenges. Other suggestions included verifying wind forecasts in spatially complex areas, such as fjords, mountainous regions, and ocean and land areas near complex coastlines, using advanced data sources like synthetic aperture radar and other remote sensing measurements.

Several forecasters also highlighted the need to verify uncertainty in probabilistic forecasts from ensemble prediction systems, and to better evaluate physical processes in the boundary layer.

While some of these aspects are already verified, forecasters noted that the results are not always communicated effectively, leaving gaps in their understanding of model performance. There was also a desire for more interaction between statistical and subjective verification approaches, e.g., case studies, as well as clearer explanations of what different verification scores mean in an operational context. These findings highlight the need for both the development of new verification methods and improved communication of verification results to better meet forecasters' operational needs.

3.4 Views on data-driven forecasts

At the time of the workshop in Tromsø, data-driven models had not yet become close to being (semi-)operational, and was therefore not a theme to discuss. However, the pace of development of data-driven models is fast and a discussion on them was included in the two later workshops. In the workshops held in Bergen and Oslo, discussions included what it would take for forecasters to trust data-driven forecasts to the same extent as traditional NWP forecasts. A recurring theme was the importance of transparent and systematic verification against both observations and NWP forecasts using standard verification methods. Forecasters considered equal or better performance compared to existing NWP systems to be an important condition for trust, particularly for parameters that are critical for operational decision-making. Verification should also cover both low intensity weather conditions and extreme events. Forecasters emphasised the importance of case studies to document how data-driven models perform in high-impact, rare and extreme conditions as these situations are crucial for building trust.

Beyond verification scores, another key concern was the need for practical experience over time. Forecasters noted that trust in any new forecasting system requires an extended test period of at least 6–12 months, to cover a wide range of different weather scenarios, when data-driven models can be tested in parallel with traditional systems. Such parallel use would allow forecasters to gain familiarity with the model behaviour through real weather situations and to identify strengths, weaknesses and possible model artefacts that may not be apparent from summary verification scores alone. To support this, it was suggested that data-driven models should be integrated to existing operational tools for direct comparison with NWP forecasts.

Transparency and understanding of how data-driven models are developed, trained and function were also viewed as critical for building trust. Forecasters expressed a desire for clear and accessible information about the datasets used for training, the quality of these datasets, the inclusion of extreme events and the physical consistency of the data-driven forecasts. They emphasised that data-driven forecasts must produce outputs that are physically realistic, synoptically plausible and internally coherent. These aspects align closely with the concept of explainability of AI-based weather prediction (Casati et al., 2026; Bröcker et al., 2026). Additionally, regular updates about improvements were seen as important for building and maintaining trust.

Finally, forecasters stressed the importance of involving them early in the development and integration of new data-driven forecasting tools. Such collaboration would ensure that forecasters' expertise is used for development and validation of these systems, and that data-driven models are aligned with operational forecaster needs.

3.5 Desired tools and knowledge

In the workshops, forecasters shared a wide range of ideas for tools and knowledge that could support their ability to assess forecasts during shifts. The recurring theme was the need for tools that provide relevant, accessible and practical information that aligns with the operational workflow. Forecasters emphasised that they do not necessarily need more data but rather better access to tailored information that can be easily used in shifts.

One request was for location-specific verification that includes recent and historical performance of key weather parameters. This information should be accessible in tools with search functionality by location, parameter and weather situation allowing forecasters to quickly retrieve verification results for specific sites. For example, one forecaster suggested tools that could provide verification of Terminal Aerodrome Forecast (TAF) elements for all airports, with options to filter by time period or regions.

Forecasters also requested summaries of recent model performance, such as maps or tables showing model biases or skill scores for specific parameters, regions or seasons. Interactive maps were suggested, highlighting areas or stations with particularly large errors and also stations with high model quality. This could include information about model mean errors separately for seasons. For extreme weather verification, forecasters suggested a situation-based reference tool that documents model strengths and weaknesses for the specific conditions.

Another need identified was near-real-time verification tools that allow forecasters to compare forecasts and observations quickly and efficiently. For example, tools that could provide information about model quality on a map or as a timeseries for the last 6 h. Forecasters also suggested tools that could automatically generate verification summaries of past forecasts that they have written and send them via email or internal chat systems. Such tools would enable continuous feedback and learning from their own work. For long-term learning and knowledge transfer, forecasters expressed interest in tools that combine historical weather events with model performance in similar situations. As one forecaster expressed it: “I wish there was a possibility to look up weather situations and how the models have previously performed in these types of cases.” This suggests a need for diagnostic verification that can be conditioned on the current weather situation, for example by synoptic weather type, wind direction or other relevant features.

Overall, the feedback highlights the need for co-development of verification tools that help forecasters interpret model quality and communicate the anticipated uncertainty introduced by known model limitations in specific forecasts.

3.6 Communication ideas

Forecasters shared several ideas for improving how knowledge about model quality could be better shared, both internally within the organisation, and externally. Internally, they emphasised the importance of centralised and accessible resources, such as an internal portal that gathers all information into one place. Suggested tools included interactive maps, the ability to search information by location or situation, and short summaries dedicated to forecasters to keep them updated. Regular opportunities for knowledge-sharing, such as case-based discussions, workshops and briefings were seen as the most effective way to learn. For example, monthly summary meetings were suggested as a way to share operational insights and model performance information across teams. Forecasters also stressed that communication should focus on practical relevance for everyday forecasting.

Externally, forecasters recommended presenting forecast quality information in a way that is understandable and relevant for various users. For the general public, simple and understandable formats such as social media posts, websites, newsletters and articles were recommended. Meteorological phenomena should be explained in plain language, with examples of why certain weather phenomena, such as convective precipitation, are difficult to predict. For advanced users, detailed verification data could be made available through dedicated platforms. Across both internal and external communication, it is essential to enhance understanding of how users interpret forecasts and to adapt communication to ensure accessibility and clarity.

4 Discussion

This study explored how operational forecasters at MET Norway evaluate and perceive model quality and identified conditions and tools necessary for building trust in model products. The findings show that forecasters’ assessments based on accumulated knowledge and experience over time align closely with known model characteristics but they also highlight needs for tailored verification tools to support operations that are not addressed fully by current verification practices.

The alignment between forecasters’ evaluations and known model limitations reflects a shared understanding of model behaviour within the institute. At the same time, the workshops revealed additional insights into model performance beyond standard station-based quantitative verification, particularly for phenomena that are less well captured by conventional station-based verification, such as fog and low clouds. These phenomena have a high operational relevance but remain challenging to evaluate using traditional metrics. This finding supports Kain et al. (2003) who demonstrated the importance of collaboration between forecasters and model developers, and emphasized that subjective evaluation and user experience provide added value that quantitative verification alone may miss. Forecasters primarily build knowledge about model quality through repeated model use during shifts. Their evaluation is based on experience-based heuristics, and reflects accumulated operational knowledge of what Doswell (2004) described as "forecaster intuition". These results suggest that experience-based evaluations by forecasters not only confirm but also complement quantitative verification. This study thus documents both forecasters' shared understanding of model quality, and the processes through which they build this understanding at a national weather service.

A central theme in the workshops was the need for verification that reflects operational situations, such as specific events, similar past events or regionally relevant phenomena. This became evident when forecasters discussed their perceived model quality, their views on data-driven models and their desired tools and knowledge. Overall, forecasters expressed less interest in aggregated summary statistics and more interest in situation-dependent evaluations. The findings also suggest that for the parameters and processes that forecasters have lower confidence in the model forecasts, verification needs to improve, and the results should be communicated more effectively to the operational forecasters. This perspective aligns with Demuth et al. (2020) who emphasised that verification information helps forecasters to judge how much to rely on different models. Our findings suggest that forecasters primarily view verification as decision support during shifts. This reframes verification as an integral part of operational practice rather than as only something that is done afterwards for reporting and development purposes. Similar to how Murphy (1993) argues about forecasts, it can be argued that verification first gets a value when used in decision-making. This is consistent with the verification information value ladder proposed by Pagano et al. (2024), which describes the need to transform verification data into actionable information and insights that support user decisions. For forecasters, this emphasises that verification needs to be tailored towards operational needs, effectively communicated to, and understood by, operational forecasters. At the same time, other types of verification, such as summary statistics, remain important for users with different priorities, such as model developers or administrators focused on monitoring or strategic goals.

The workshops reveal that forecasters use both experience-based and quantitative verification to assess when model guidance can be trusted. This was particularly evident in discussions about new data-driven forecasts. Padilla et al. (2021) demonstrated that even qualitative descriptions of confidence influences trust in forecasts and decision-making. Similarly, Fundel et al. (2019) emphasised that hands-on use and real-time experience builds understanding and trust on new model guidance. This study extends the discussion from end-users to operational forecasters, and shows that trust is not only about objective model quality, but also about experience, transparency, and availability of suitable verification metrics. Trust appears to be a dynamic process that is built through repeated assessment of model quality across various weather situations. By linking verification directly to trust-building in operational forecasting, this study provides valuable insights into how forecasters integrate new types of model guidance into their work.

Forecasters identified concrete, operationally relevant needs for tools that reflect operational practice, including event-based verification, filtering by phenomenon, flow pattern or region, and interactive approaches. These needs are consistent with verification research, such as conditional and process-based verification (Casati et al., 2008). Previous studies have shown that traditional verification methods, such as grid-box averages compared to station observations can mislead the perceived model performance (Göber et al., 2008). Observational uncertainties can further complicate verification results, with undercatch of solid precipitation by gauges (Køltzow et al., 2020) being particularly relevant to Norway, especially in the winter. Forecasters reported local challenges, such as systematic model errors in specific locations, which can partly be attributed to representativeness issues, and may reduce trust in model outputs. To address these issues, verification tools should address the methodological challenges. In addition, they should be integrated into forecasters' near-real time workflow and enable quick and reliable decision-making, especially under strict time constraints of issuing weather warnings, when rapid interpretation of model outputs and uncertainties is essential.

The workshops highlight the value of structured dialogue between model developers and operational forecasters. Similar initiatives, such as the Spring Programme in the USA described by Kain et al. (2003), have demonstrated the benefits of such collaborations. Steenburgh et al. (2013) emphasised that bridging research and operations requires effective communication and mutual understanding. They also supported the idea of co-developing tools, fostering mutual understanding between researchers and forecasters, and integrating research into operational practices. This study provides an example of how forecaster-developer collaboration can be initiated within a national weather service, offering a model that could be transferable to other institutes. By creating space for structured dialogue, the workshop revealed new perspectives, previously hidden assumptions, and concrete needs, which can shape how verification and its communication can be approached in the future.

This study was conducted within a single national weather service and reflects the specific context of MET Norway. As forecasting practices, verification systems and organisational structures differ between countries, many of the findings may offer insights that generalize to similar operational settings. Data were collected in an informal workshop setting using facilitator notes and short written responses. While this approach enabled broad participation and open discussion in a safe environment, it lacked the depth of structured qualitative interviews. Consequently, the findings represent descriptive themes and may have been influenced by the framing provided by the model developers and verification scientists who served as facilitators.

The findings provide a basis for further development of verification tools and communication practices tailored to operational needs. This includes improving existing quantitative verification to make it more relevant for decision-making during shifts and developing tools that support forecasters' subjective learning processes about model quality. For example, access to past cases, event-based verification and situation-dependent summaries could help forecasters build experience more efficiently. Future work could also include the co-design of verification products in collaboration with forecasters. Additionally, more detailed qualitative studies could provide deeper insight into how verification is applied in practice and how forecasters build trust in model guidance. Such studies could investigate more systematically which factors shape this trust, including situations where trust may be over- or underestimated in relation to independent verification evidence. Follow-up workshops could focus more explicitly on the operational use of ensemble forecasts and probabilistic information. Data-driven forecasts will also need to be addressed when their use becomes more established in operations. Finally, as forecaster feedback is already collected and shared between Norway, Sweden, Finland and Estonia, extending the collaboration to cross-institutional comparisons between national weather centres would help assess the transferability of the findings and identify common challenges and solutions.

5 Conclusions

This study aimed to understand forecasters' perspectives on model quality, how they gather and apply information on model quality in their work, and what tools would help them build trust in model guidance. To achieve this, a series of workshops was conducted at the three main forecasting offices in Norway.

The main findings are as follows:

- Forecasters' experience-based evaluation confirms and complements quantitative verification.
- Forecasters expressed a need for verification methods that reflect operational situations, such as specific events or similar past events.
- Both experience-based and quantitative verification are used by forecasters to assess when models can be trusted.
- Forecaster input should contribute to the co-development of verification methods and tools, and communication of them.
- Structured workshops can support dialogue between forecasters and model developers and help integrate forecaster perspectives into verification and system development.

These findings indicate a need for verification approaches that are tailored to the needs of operational forecasters and support both decision-making during shifts and trust-building in model guidance. More broadly, they show how verification and evaluation can support operational forecasting by making model performance information more relevant for forecasters and by helping model development focus on user priorities.

The next steps at MET Norway include expanding the operational verification and monitoring system to cover more user-requested variables, incorporating conditional verification and further tailoring verification tools to meet the forecasters' needs. Future research should explore forecasters' perspectives in more depth, particularly regarding the use of ensemble forecasts and data-driven models. Finally, fostering structured collaboration between forecasters and model developers is essential for bridging communication gaps and enabling meaningful dialogue that benefits both communities.

Code and data availability. The aggregated results needed to support the findings are presented in the article. The raw workshop material is not publicly available because consent was not obtained for public archiving and individual participants could be indirectly identifiable. The scripts used to plot the figures are available upon reasonable request.

Author contributions. MK, LØ, AS and TR initiated the workshops, GN, MSS, and BGL managed the local workshop arrangements, MK, LØ, and TR collected facilitator notes, AS and TR prepared the figures, and TR coordinated the writing process and prepared the initial manuscript draft. All authors discussed the results and contributed to the review and editing of the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Special issue statement. This article is part of the special issue "EMS Annual Meeting: European Conference for Applied Mete-

orology and Climatology 2025”. It is a result of the EMS Annual Meeting 2025, Ljubljana, Slovenia, 7–12 September 2025. The corresponding presentation was part of session ES2.2: Communicating science and dealing with Uncertainties.

Acknowledgements. We thank participants of the workshops from the forecasting division for their valuable contributions. We also acknowledge Eivind Støylen and Rafael Grote for contributing to the initial idea and their involvement in the workshop in Tromsø.

Review statement. This paper was edited by Gerald Fleming and reviewed by Thomas Pagano, Barbara Casati, and one anonymous referee.

References

Ainslie, B., So, R., and Chen, J.: Operational Evaluation of a Wildfire Air Quality Model from a Forecaster Point of View, Weather Forecast., 37, 681–698, https://doi.org/10.1175/WAF-D-21-0064.1, 2022.

Bouallegue, Z. B., Haiden, T., Weber, N. J., Hamill, T. M., and Richardson, D. S.: Accounting for Representativeness in the Verification of Ensemble Precipitation Forecasts, Mon. Weather Rev., 148, 2049–2062, https://doi.org/10.1175/MWR-D-19-0323.1, 2020.

Bröcker, J., Driscoll, S., Necker, T., Rodríguez, J., Dacre, H., Harvey, N., and Ben Bouallègue, Z.: Verification of AI-based environmental forecasting systems: What can we do, what do we need to do, and what are the challenges?, J. Eur. Meteorol. Soc., 4, 100032, https://doi.org/10.1016/j.jemets.2026.100032, 2026.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S.: Forecast verification: current status and future directions, Meteorol. Appl., 15, 3–18, https://doi.org/10.1002/met.52, 2008.

Casati, B., Loveday, N., Bouallègue, Z. B., Bröcker, J., Wahl, S., de Elía, R., Pic, R., Subich, C., Separovic, L., Brunet, D., and McTaggart-Cowan, R.: Operational Verification of Artificial Intelligence Weather Prediction, B. Am. Meteorol. Soc., BAMS–D–26–0054.1, https://doi.org/10.1175/BAMS-D-26-0054.1, 2026.

Demuth, J., Morss, R., Jankov, I., Alcott, T., Alexander, C., Nietfeld, D., Jensen, T., Novak, D., and Benjamin, S.: Recommendations for Developing Useful and Usable Convection-Allowing Model Ensemble Information for NWS Forecasters, Weather Forecast., 35, 1381–1406, https://doi.org/10.1175/WAF-D-19-0108.1, 2020.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L. J.: The Setup of the MesoVICT Project, B. Am. Meteorol. Soc., 99, 1887–1906, https://doi.org/10.1175/BAMS-D-17-0164.1, 2018.

Doswell III, C.: Weather Forecasting by Humansâ€”Heuristics and Decision Making, Weather Forecast., 19, 1115–1126, https://doi.org/10.1175/WAF-821.1, 2004.

ECMWF: IFS Documentation CY49R1, https://www.ecmwf.int/en/publications/ifs-documentation (last access: 4 August 2026), 2024.

Eresmaa, R., Andrae, U., Berggren, L., Bremnes, J. B., Fortelius, C., Frogner, I., Gregow, E., Grote, R., Ridal, M., Vignes, O., Ansper, I., Azad, R., Begens, J., Bergholt, L., Blyverket, J., Daniel, L., Engdahl, B. J., Erlandsen, H., Eshagh, M., Hasu, M., Homleid, M., Ideström, P., Ivarsson, K., Lerner-Vilu, A., Mets, A., Moldengauere, S., Noer, G., Nurmi, E., Partio, M., Priedītis, G., Saarikalle, E., Schönach, D., Shapkalijevski, M., Sild, K., SitĊihins, V., Sokka, N., Spjelkavik, S., Tack, A., Vaštšenko, A., Wettergren, A., Ylinen, K., and Zandovskis, U.: The Operational Forecast Process at MetCoOp, B. Am. Meteorol. Soc., 107, E946–E963, https://doi.org/10.1175/BAMS-D-25-0225.1, 2026.

Frogner, I.-L., Singleton, A. T., Køltzow, M., and Andrae, U.: Convection-permitting ensembles: Challenges related to their design and use, Q. J. Roy. Meteor. Soc., 145, 90–106, https://doi.org/10.1002/qj.3525, 2019.

Fundel, V. J., Fleischhut, N., Herzog, S. M., Göber, M., and Hagedorn, R.: Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users, Q. J. Roy. Meteor. Soc., 145, 210–231, https://doi.org/10.1002/qj.3482, 2019.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spatial Forecast Verification Methods, Weather Forecast., 24, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1, 2009.

Göber, M., Zsótér, E., and Richardson, D. S.: Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification, Meteorol. Appl., 15, 359–365, https://doi.org/10.1002/met.78, 2008.

Kain, J., Janish, P., Weiss, S., Baldwin, M., Schneider, R., and Brooks, H.: Collaboration between Forecasters and Research Scientists at the NSSL and SPC: The Spring Program, B. Am. Meteorol. Soc., 84, 1797–1806, https://doi.org/10.1175/BAMS-84-12-1797, 2003.

Køltzow, M., Casati, B., Bazile, E., Haiden, T., and Valkonen, T.: An NWP Model Intercomparison of Surface Weather Parameters in the European Arctic during the Year of Polar Prediction Special Observing Period Northern Hemisphere 1, Weather Forecast., 34, 959–983, https://doi.org/10.1175/WAF-D-19-0003.1, 2019.

Køltzow, M., Casati, B., Haiden, T., and Valkonen, T.: Verification of Solid Precipitation Forecasts from Numerical Weather Prediction Models in Norway, Weather Forecast., 35, 2279–2292, https://doi.org/10.1175/WAF-D-20-0060.1, 2020.

Køltzow, M., Schyberg, H., Støylen, E., and Yang, X.: Value of the Copernicus Arctic Regional Reanalysis (CARRA) in representing near-surface temperature and wind speed in the north-east European Arctic, Polar Res., 41, https://doi.org/10.33265/polar.v41.8002, 2022.

Køltzow, M., Øestad, M., Azad, R., Rugaard Furevik, B., Olsen, A.-M., and Remes, T.: How inaccurate offshore wind observations impact the quality of operational numerical weather prediction, Q. J. Roy. Meteor. Soc., 150, 5337–5355, https://doi.org/10.1002/qj.4872, 2024.

Lillegraven, B. G., Noer, G., and Østvand, L.: Verification of Operational Weather Prediction Models: September to November 2024, METinfo 21/2025, Norwegian Meteorological Institute, ISSN 1894-759X, https://www.met.no/publikasjoner/met-info/ (last access: 4 August 2026), 2025a.

Lillegraven, B. G., Noer, G., and Østvand, L.: Verification of Operational Weather Prediction Models: December 2024 to February

2025, METinfo 31/2025, Norwegian Meteorological Institute, ISSN 1894-759X, https://www.met.no/publikasjoner/met-info/ (last access: 4 August 2026), 2025b.

Lillegraven, B. G., Noer, G., and Østvand, L.: Verification of Operational Weather Prediction Models: March to May 2025, METinfo 36/2025, Norwegian Meteorological Institute, ISSN 1894-759X, https://www.met.no/publikasjoner/met-info/ (last access: 4 August 2026), 2025c.

Lillegraven, B. G., Noer, G., and Østvand, L.: Verification of Operational Weather Prediction Models: September to November 2025, METinfo 18/2026, Norwegian Meteorological Institute, ISSN 1894-759X, https://www.met.no/publikasjoner/met-info/ (last access: 4 August 2026), 2026.

Müller, M., Batrak, Y., Kristiansen, J., Køltzow, M. A., Noer, G., and Korosov, A.: Characteristics of a Convective-Scale Weather Forecasting System for the European Arctic, Mon. Weather Rev., 145, 4771–4787, https://doi.org/10.1175/MWR-D-17-0194.1, 2017.

Murphy, A.: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, Weather Forecast., 8, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, 1993.

Padilla, L. M. K., Powell, M., Kay, M., and Hullman, J.: Uncertain About Uncertainty: How Qualitative Expressions of Forecaster Confidence Impact Decision-Making With Uncertainty Visualizations, Front. Psychol., 11, 579267, https://doi.org/10.3389/fpsyg.2020.579267, 2021.

Pagano, T. C., Casati, B., Landman, S., Loveday, N., Taggart, R., Ebert, E. E., Khanarmuei, M., Jensen, T. L., Mittermaier, M., Roberts, H., Willington, S., Roberts, N., Sowko, M., Strassberg, G., Kluepfel, C., Bullock, T. A., Turner, D. D., Pappenberger, F., Osborne, N., and Noble, C.: Challenges of Operational Weather Forecast Verification and Evaluation, B. Am. Meteorol. Soc., 105, E789–E802, https://doi.org/10.1175/BAMS-D-22-0257.1, 2024.

Pagano, T. C., Ebert, E. E., and Khanarmuei, M.: Enhancing Forecast Verification in National Meteorological and Hydrological Services, Meteorol. Appl., 32, e70051, https://doi.org/10.1002/met.70051, 2025.

Remes, T., Køltzow, M., and Kähnert, M.: Spatial variability of near-surface air temperature in the Copernicus Arctic regional reanalysis, Q. J. Roy. Meteor. Soc., 151, e4941, https://doi.org/10.1002/qj.4941, 2025.

Steenburgh, W. J., Schultz, D. M., Snyder, B. J., and Meyers, M. P.: Bridging the Gap Between Operations and Research to Improve Weather Prediction in Mountainous Regions, in: Mountain Weather Research and Forecasting, Springer, https://doi.org/10.1007/978-94-007-4098-3_12, 2013.

Vigna, I., Jeuring, J., and Doksæter Sivle, A.: Melkeværet: A serious game to bridge climate service providers and agricultural users, Tech. Rep. MET report 16/2024, Norwegian Meteorological Institute, https://www.met.no/publikasjoner/met-report/met-report-2024 (last access: 4 August 2026), 2024.