BAMS Article

Challenges of Operational Weather Forecast Verification and Evaluation

Thomas C. Pagano,[a] Barbara Casati,[b] Stephanie Landman,[c] Nicholas Loveday,[a] Robert Taggart,[a] Elizabeth E. Ebert,[a] Mohammadreza Khanarmuei,[a] Tara L. Jensen,[d,e] Marion Mittermaier,[f] Helen Roberts,[f] Steve Willington,[f] Nigel Roberts,[f] Mike Sowko,[g] Gordon Strassberg,[g] Charles Kluepfel,[g] Timothy A. Bullock,[h] David D. Turner,[i] Florian Pappenberger,[j] Neal Osborne,[k] and Chris Noble[k]

KEYWORDS: Uncertainty; Forecast verification/skill; Numerical weather prediction/forecasting; Operational forecasting; Communications/decision-making

ABSTRACT: Operational agencies face significant challenges related to the verification and evaluation of weather forecasts. These challenges were investigated in a series of online workshops and polls engaging operational personnel from six countries. Five key themes emerged: inadequate verification approaches for both existing and emerging products; incomplete and uncertain observations; difficulties in accurately capturing users' real-world experiences using simplified metrics; poor communication and understanding of forecasts and complex verification information; and institutional factors such as limited resources, evolving meteorologist roles, and concerns over reputational damage. We identify nearly 50 operationally relevant scientific questions and suggest calls to action. Addressing these needs includes designing forecast systems with verification as a central consideration, enhancing the availability of observations, and developing and adopting community software systems. Additionally, we propose the establishment of an international community comprising environmental and social science researchers, statisticians, verification practitioners, and users to provide sustained support for this collective endeavor.

DOI: 10.1175/BAMS-D-22-0257.1

Corresponding author: Thomas C. Pagano, tom.pagano@bom.gov.au

Manuscript received 22 November 2022, in final form 21 January 2024, accepted 26 January 2024

AFFILIATIONS: [a] Bureau of Meteorology, Melbourne, Victoria, Australia; [b] Environment and Climate Change Canada, Dorval, Quebec, Canada; [c] South African Weather Service, Pretoria, South Africa; [d] National Center for Atmospheric Research, Boulder, Colorado; [e] Developmental Testbed Center, Boulder, Colorado; [f] Met Office, Exeter, United Kingdom; [g] National Weather Service, Silver Spring, Maryland; [h] Environment and Climate Change Canada, Dartmouth, Nova Scotia, Canada; [i] NOAA/Global Systems Laboratory, Boulder, Colorado; [j] European Centre for Medium Range Weather Forecasts, Reading, United Kingdom; [k] Meteorological Service of New Zealand, Wellington, New Zealand

1. Introduction

Weather prediction (Fig. 1) is a continuously evolving enterprise. The rapid expansion of various types of data (including satellite and other kinds), coupled with increasing computer capabilities, advances in numerical models, and sophisticated postprocessing, has led to a steady increase in forecast accuracy. Forecast skill in the range from 3 to 10 days ahead has been increasing by about 1 day per decade (Bauer et al. 2015). Better forecasts benefit the public in their daily lives and are also essential for various industries, emergency management communities, and governments (World Meteorological Organization 2015).

Nevertheless, operational agencies face growing pressure to demonstrate value and accountability, particularly in the aftermath of extreme weather, shifting funding priorities, and competition from other forecast providers. To address these challenges and drive improvements, agencies must evaluate performance along every link of the value chain (Golding 2022, as depicted in Fig. 2).

Verification, in this context, is the process of comparing forecasts with observations or their proxies, to assess the quality and value of the forecasts (Murphy 1993). It is an essential component of performance assessment, which encompasses additional factors such as timeliness, consistency, relevance, accessibility, and interpretability but also the efficiency of the forecasting service (World Meteorological Organization 2000). Some authors use the terms verification and evaluation interchangeably, although, in our context, evaluation is a more holistic investigation of how and why forecasts differed from observations. "Operational" refers to forecasts produced by an institution (typically a government agency but sometimes a private company) providing an ongoing and supported time-critical service. Our primary focus is on core predictive services, such as routine publicly available weather forecasts, public warnings, and other environmental predictions (e.g., fires or floods). While operational agencies may have in-house capabilities for hindcasts, analyses, and numerical weather prediction (NWP) model development research, we emphasize these areas less.

Fig. 1. Operational meteorologists from the South African Weather Service reviewing recent weather conditions.

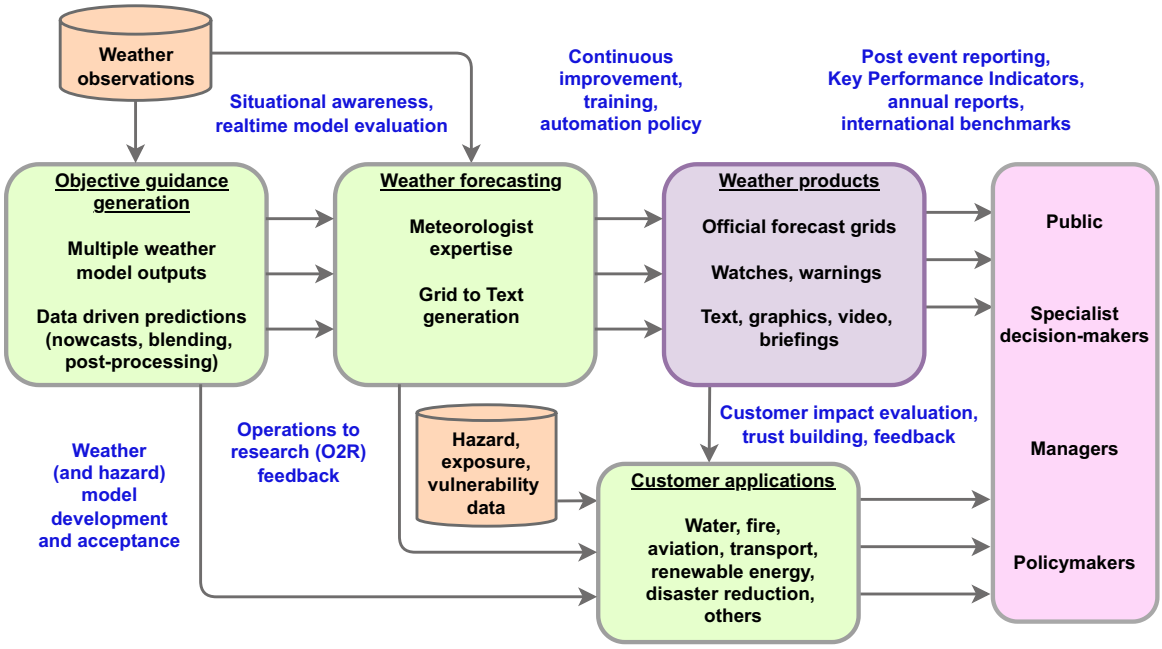

Fig. 2. Schematic view of interconnections in the weather forecasting value chain. The blue text shows the verification and evaluation-related activities.

To support these operational activities, a rapidly growing community continues to push the boundaries of verification research (Casati et al. 2008, 2022; Ebert et al. 2013). Recent technological breakthroughs have enabled users to access verification tools, software, and information in unprecedented ways (such as Brown et al. 2021). While some weather modeling centers have highly advanced verification systems, some countries' operational practices are lagging the state of the science, particularly when it comes to supporting users (Casati et al. 2008).

Most academics struggle to know the internal challenges encountered by operational agencies. Documentation of operational practices and concerns, particularly in developing and least-developed countries, can be scarce (Casati et al. 2008). Even within and among agencies, there can be confusion about who assesses the quality of specific products along the value chain, how they do it, and what they do with the results.

This article aims to give a unified voice to forecast producers to help overcome these challenges. The authors organized online workshops in 2023, bringing together over 50 operational personnel from six countries, including model developers, observation managers, operational meteorologists, and verification system developers. Through discussions and polls, specific challenges were identified, and potential solutions proposed.

We discuss five categories of challenges (summarized in Table 1), list important science questions relevant to each, and conclude with calls to action as a collaborative community.

2. Operational challenges for weather services

a. Challenge 1: Verification approaches. In 1951, a panel discussed the challenges of operational weather forecast verification and stated "Unfortunately, much time has been wasted on attempts to devise a method or single score that will serve all purposes. The very nature of weather forecasts and verifications and the way they are used make one single ... standard of evaluation impossible" (Allen et al. 1952). Decades later, serving the breadth of interests (Table 2) remains the primary operational challenge. While general performance statements, such as "90% of 1-day-ahead temperature forecasts were within 2°C from the observed," may be interesting, personalized information specific to locations, lead times,

Table 1. Verification and evaluation challenges faced by operational agencies.

Verification approaches	1) No single approach meets every user's needs 2) Some legacy verification methods are inadequate for emerging forecasts 3) Some important forecasts are unstructured/qualitative, hindering objective verification 4) Unprecedented weather situations are increasingly common
Observations	1) Some observations are of low quality, and some are good but unrepresentative 2) Networks can be sparse and data incomplete 3) Unconventional observations are underutilized 4) Databases of impacts are lacking or restricted
Evaluation of forecast impact	1) "When it matters most" can be a very small sample size 2) Impact-based verification can be confounded by user responses to warnings 3) Models of users' decision-making processes have many assumptions 4) Verification along the value chain is resource intensive
Communication	1) Some forecasts deserve partial credit for being close 2) Verification jargon and concepts are unfamiliar 3) Evaluation usually ignores common forecast misinterpretations 4) Intermediaries to interpret verification results are rare
Institutional factors	1) Agencies fear reputational damage or loss of support 2) The role of humans is changing through automation 3) Operations-to-research feedback opportunities are evolving 4) Some agencies are highly constrained

seasons, and weather parameters of interest is more useful (Hartmann et al. 2002). Tailoring such information requires significant effort.

As agencies grapple with verifying traditional forecasting products, they also need to provide new types of verification for emerging products. Traditional approaches focus on single-valued or categorical forecasts, while forecasts have been evolving to increasingly rely on probabilities via ensembles, statistical approaches, or other techniques. While ensemble and probabilistic verification approaches tend to be statistically rigorous and proper (Gneiting and Katzfuss 2014), many of their outputs are not intuitive, and meteorologists are concerned that familiar tools for situational awareness and model diagnosis may become obsolete without suitable replacements. Furthermore, agencies are expanding their scope to include rapid update models; hazard- and risk-based forecasts, such as fire, floods, and thunderstorm asthma epidemics (Bannister et al. 2021); and forecasting features like sea ice edge (Palerme et al. 2019) or the onset of the rainy season: these emerging forecasts often need to harvest specific observations and use specialized verification approaches,

Table 2. Examples of verification and evaluation information needs of various stakeholders.

Model developers and postprocessors	1) Are there patterns (trends, cycles, and episodes) to the errors? What causes them? 2) Does the forecast quality and its changes over time (both year to year and across different lead times) match our expectations? 3) Does the upgraded/postprocessed system outperform or behave differently from the original? 4) What system improvements should I prioritize?
Meteorologists and forecast users	1) How much should I trust today's forecast or source of objective guidance? Is this different from yesterday? 2) As a meteorologist, how can I add value to this forecast? 3) Can I trust this forecast to provide information on the full range of possible outcomes, including the reasonable worst-case scenario? 4) What can I learn from recent performance? 5) Is the forecast certainty compatible with my risk tolerance?
Managers and administrators	1) What is the overall quality of the service we are providing? 2) How are we performing compared to our competitors? 3) Have we realized the benefits of past investments? 4) Were our forecasts of high-impact events useful to decision-makers?

which are resource demanding. A good example of the latter is convective-scale NWP verification using spatial methods that measure forecast realism without requiring precise agreements at points.

Public weather forecasts often take the form of semistructured text or icons. Some of these are generated automatically; for example, Australian weather symbols contain raindrops if the chance of daily precipitation is at least 35%. This threshold is not widely known, and it is an open question if the verification scores reflect the users' true satisfaction, given that a seemingly "perfectly reliable" forecast for raindrops may end up dry 65% of the time. Similarly, when confronted with vague phrases like "possible storms," individuals' interpretations vary, mostly ranging between 10% and 50% chance (Lenhardt et al. 2020). Certain icons (such as dust, haze, and frost) lack direct observations for comparison.

Some agencies also place emphasis on generating a "policy" or "advice" product, a narrative of "the weather story" for briefing stakeholders (Fig. 3). During extreme events, face-to-face consultation between emergency managers and embedded meteorologists may end up much more impactful than the numbers within the official forecasts. However, narrative-style forecasts are difficult to deconstruct and process objectively. The content may have an inconsistent structure from one product issue to the next. They occasionally omit certain pieces of information leading to many "missing" forecasts, thus impacting those verification scores which assume complete data. Many of these essential products can only be evaluated qualitatively (if at all).

Increasingly, a nonstationary climate is revealing the limitations of some statistical and physics-based models. For example, during the recent U.K. record-breaking heatwave, some NWP models produced improbably high temperature forecasts because of an issue with overdrying of modeled soils, which was exposed when the model was pushed into uncharted territory. Meteorologists quickly consulted model developers and intervened to prevent a major overforecast bust. On the other hand, during a recent record low pressure system and cold snap in the United States, researchers and forecasters questioned the plausibility of the model predictions, but the event still occurred. Attribution studies aim to assess the extent to which climate change is responsible for specific extreme weather events: to what extent can forecast inaccuracies be attributed to poorly modeled processes that become increasingly important in a changing climate?

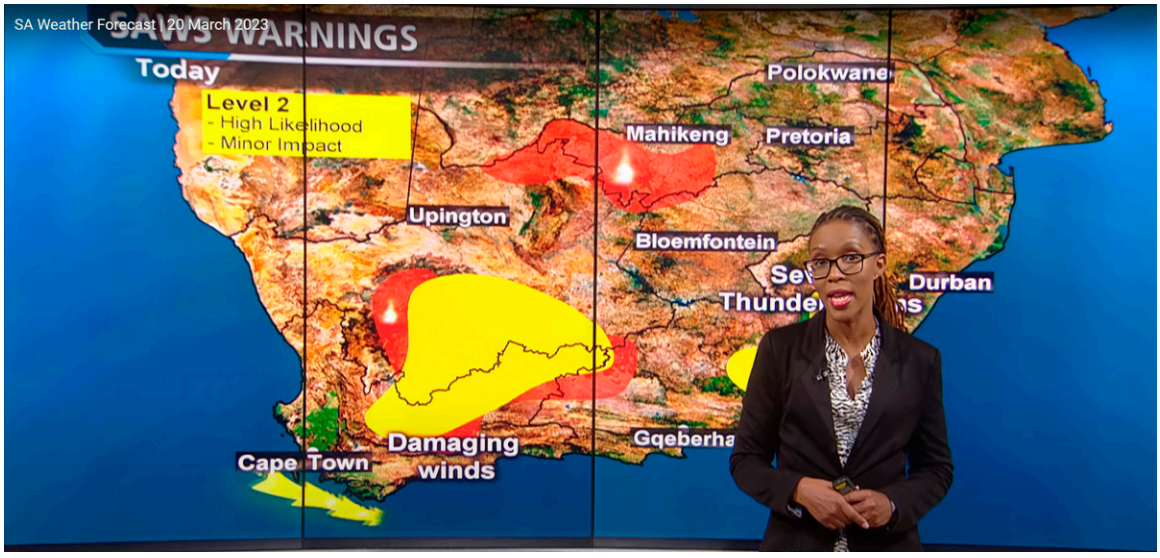

Fig. 3. A South African Weather Service meteorologist delivers a weather briefing containing statements of impacts. How can such "story-based" products be verified? https://youtu.be/ZQupmOd97Jw?t=55.

Research opportunities. How can long operationally realistic hindcast datasets be generated more efficiently (e.g., reduced ensembles) to provide a larger set of verification cases through similar computing effort? Can automated language models be used to disassemble text-based forecasts for use in verification? How can products like the weather story be verified objectively? How to best verify short-range forecasts whose accuracy is comparable to observational uncertainty? What are the best user-oriented approaches for verifying seamless forecasts, where both the user needs and scales of useful information vary with lead time? Can we project how climate change will affect forecast accuracy? Are the past effects of climate change detectable in multidecadal verification time series? Can spatial verification methods be extended to emphasize predictive performance for extremes without creating forecaster's dilemmas (where the forecaster does not predict what they "really think is going to happen" because a different strategy promises better performance scores, Taggart 2022; Lerch et al. 2017)? How to better integrate spatial and novel verification methods into operational practices? How to use observations to provide insights into poorly modeled processes?

b. Challenge 2: Observations. In weather forecasting, a major hurdle arises from the inherent limitations of observations, leading to an incomplete picture of the atmospheric conditions (Fig. 4). The challenge lies in the fact that various factors, such as nonstationary and systematic errors (e.g., sensor drift or errors in how the data are processed and interpreted), can impact the accuracy of measurements. Additionally, it is often difficult to ascertain whether local weather observations accurately represent a larger geographic area or a specific model grid cell. While it has long been proposed that observations should be treated probabilistically, operational practice in verification is to handle them deterministically in nearly all cases.

Data assimilation routinely censors or discounts problematic observations, but each mismatch between analysis and observation tells a potentially interesting story. Remote sensing data and model-generated analyses offer good spatial coverage but come with their own

Fig. 4. The Wattamolla Automatic Weather Station in Australia had a bent wind sensor in 2020, impacting measurement accuracy. Even under normal conditions, the station's data do not fully reflect the weather on land or at sea due to its proximity to a 180-ft-tall cliff.

quality issues, including inheriting model biases, with uncertainties sometimes comparable to forecast errors.

Information about impacts is often much less available than weather observations. Manually scouring social media for damage reports is labor intensive, yielding patchy and inconsistent results (such as a bias toward urban areas). Some datasets, including weather observations, may be restricted or commercially sensitive. Many long-standing hazard-based products lack direct observations for comparison. For example, fire danger indices reflect the potential effort required to contain a runaway blaze. How does one directly assess the danger of fires that never began? Instead, agencies sometimes use a proxy to assess impact, such as a (decades old) model relating weather parameters to fire danger.

Research opportunities. How to best account for uncertainties caused by data sparsity, gridding procedures, or the use of proxy observations? How to identify nonrepresentative stations independently of forecasts, such as using station metadata or characteristics of the observations (e.g., highly sheltered wind stations have more turbulent wind speeds)? How to handle observation certainty and analysis model dependence in scoring rules and interpretation of verification results? How to differentiate observation uncertainties and representativeness from model and forecast errors? How to verify forecasts with high resolution (e.g., 100 m) or over complex terrain (viz., what observations should be used)? How can verification approaches better accommodate temporarily missing data, especially spatial methods? Are there efficient and scalable methods and standards for collecting hazard and impact data? Which new observations enable user-oriented and high-impact verification (Marsigli et al. 2021)? Can machine learning be used to quality control traditional and novel observations? How should different observational datasets be blended, and should the approach depend on the verification aim?

c. Challenge 3: Evaluation of forecast impact. There is a heavy emphasis on evaluating performance "when it matters most" and this often gets lost in bulk scores. Verification requires a large collection of independent events to draw confident conclusions. But when it matters most is often during extreme/high-impact events and postevent reports analyze individual cases, leading to potential selection bias. For example, observed high-impact severe weather events are closely scrutinized, while false alarms are commonly ignored. Nonetheless, case studies and extremes-focused verification are often trailheads to greater insights and their ability to complement bulk scores of standard weather is evident. For instance, Lupo et al. (2023) conducted a comprehensive 7-yr study of model results, confirming operational meteorologists' impressions of the U.S. weather model's tendency to progress specific types of weather features too rapidly downstream.

Effective forecasts can also mitigate impacts, leading to fewer verifiable events. For example, when pilots are directed away from turbulence or flights are canceled in anticipation of snow, there are fewer pilot reports available to verify the warnings. However, numerous apparent false alarms do not necessarily indicate warning system failures. It is not feasible, though, to have a control group from which warnings are withheld.

To justify expenditures and investments, there is increasing pressure to frame effectiveness in financial terms. For example, an economic study of the High-Resolution Rapid Refresh (HRRR) model demonstrated that improvements would save tens to hundreds of millions of USD for various sectors, primarily wind power generation (Turner et al. 2022). Such reports depend heavily on assumptions (such as the weather sensitivity and responsiveness of various economic sectors) that are difficult to verify.

While there is a vast research literature on framing verification in terms of user benefit, most techniques assume a theoretical optimal decision-maker who reacts predictably to

the forecasts. With perhaps the exception of highly optimized renewable energy providers, there is overwhelming evidence that users' decisions are varied and context dependent (Rodwell et al. 2020) and frequently include heuristics and biases (Losee et al. 2017). When asked to describe their decision-making processes, many users would struggle to quantify their decision thresholds or forecast-related costs and losses, especially when lives are at stake (Fundel et al. 2019). Even when the user decision-making process is knowable and not commercially confidential, agencies may only have the resources to engage a select few "power" users (which raises equity issues).

Forecasts themselves are the result of a complex chain of influences. Some agencies primarily verify "upstream" products like the NWP model output, even though users do not always directly receive this information. The performance of each step in this science-to-service chain (Golding 2022) should be verified individually, although this compounds the verification effort.

In practice, making decisions based on verification information can be difficult because the results are rarely definitive. Model developers run hindcast experiments and it is common that some aspects of the forecast get better, and others get worse. Although verification is based on limited samples, hypothesis tests or confidence intervals are often not used to assess whether differences in forecast performance are statistically significant (even though approaches for doing so are relatively straightforward in most cases). A model upgrade may show, on balance, a reduction in overall temperature bias, but this masks a cold bias in the mornings that would negatively impact the timing of convective rainfall. While temperature bias is an easy headline score to track, a nuanced process-oriented approach is necessary to investigate important aspects of forecast realism. What the user values will determine the course of action in the face of such trade-offs.

Research opportunities. How can verification effectively contribute to postevent debriefings in terms of content and communication? Can causal inference assess the effectiveness of risk- and impact-based forecasts when users take mitigating actions? Can we separate the effects of changing vulnerability and exposure from changing forecast quality when investigating impact trends? How reliable are economic models in assessing the value of weather services? In terms of quality and credibility, should economic evaluation be conducted within weather services, by users, or by independent experts? Which verification approaches best align with users' perspectives without assuming an optimal decision-maker? Can machine learning be used to replicate the public's response and users' decisions toward weather warnings, enabling weather agencies to customize the warnings to minimize weather impacts? What are user perspectives on spatial displacement errors (Sherman-Morris et al. 2022)?

d. Challenge 4: Communication. The ultimate value of forecasts is in their ability to positively influence users' decisions (Murphy 1993). While verification assesses how well the forecasts match the observations, this sometimes does not tell the full story of the user's experience. There are times in which forecasts might deserve partial credit. Established methods exist for accounting for near misses in space and time (Gilleland et al. 2009) which will become all the more important for ultra-fine-resolution forecasts. However, consider a scenario where the weather icon predicts "heavy rain tomorrow" and that rain comes when most people are sleeping. Some users may feel this was a good forecast and others may feel it was a bad one (although both users would likely have benefited from digging into the more specific subdaily forecasts that most agencies now provide). Specialized airport forecasts simultaneously predict many weather elements that relate to dangerous plane landing conditions. A forecaster may feel "by the numbers, I got most of the forecast wrong, but the important parts were right." This complexity influences the

acceptance of fully automated forecasting systems solely evaluated based on quantitative metrics rather than societal outcomes.

Even when verification reflects the experiences of specific users, they might be unaware of or hesitant to accept the results due to jargon and unfamiliar concepts. Verification communication is most effective when it is clear, concise, holds personal significance for the audience, uses meaningful units, and conveys a sense of physical realism (Wang and Strong 1996). Therefore, statements like "our area under the receiver operating characteristic curve is 0.83" may be meaningful to some specialists but have no resonance with many casual users of verification information. Although NWP modeling centers routinely employ a range of complex metrics, the use of inaccessible language often prompts downstream users and managers to favor simpler ad hoc scores even though some of these scores can be easily gamed.

Misunderstanding extends beyond verification as forecasts, themselves, are also frequently misunderstood (Fleischhut et al. 2020). Differing interpretations between forecasters and users have persisted for decades (Murphy et al. 1980), and this situation also includes cases where users infer information not contained in the forecasts (Joslyn and Savelli 2021). People perceive forecasts with higher probability as more accurate (e.g., an 80% chance of rain seems more accurate than a 20% chance of no rain, Ripberger et al. 2022). Communication issues are compounded when providing information in many languages (11 in the case of South Africa). Misunderstandings affect the relevance of any verification "by the numbers," and the situation is even more challenging with ineffectively communicated probabilistic forecasts.

Verification systems are built to generate myriad numbers under the assumption that their various audiences will self-serve in accessing and understanding the information. There is rarely investment in intermediaries who can curate the right information (such as selecting thresholds of interest to the user), translate it into something actionable, and communicate it meaningfully. Only a few well-resourced agencies support a real-time "model evaluation desk" to track performance, advise meteorologists, and provide operations-to-research feedback.

Research opportunities. How to design verification that addresses users' behavioral biases and misinterpretations of forecasts? How can social media and digital technologies effectively engage users in communicating forecast quality and gathering feedback? What role do case studies and anecdotes play in understanding and conveying forecast quality? What are the best approaches for communicating probabilistic/ensemble verification using rigorous scores? How to effectively teach verification concepts and interpretation of results? Can machine learning generate insightful plain-language summaries of verification datasets? What scalable methods exist for the collaborative design of verification products and services, involving stakeholders and end-users? How can social science and graphical design enhance communication of performance information, particularly through interactive displays? What can meteorology learn from medicine and other fields about successfully communicating quantitative performance information to diverse audiences? If forecasts are communicated with quantified certainty, is there still a demand for additional verification information potentially presented alongside the forecasts (Hogan Carr et al. 2021)?

e. Challenge 5: Institutional factors. Public webpages talking about forecast performance are rare, often brief, and sometimes promotional. They present highly oversummarized statistics with limited relevance to any individual user. Organizations often are reluctant to expose faults for fear of liability or reputational damage, exposing the agency to (potentially unjust) criticism from customers or competitors. Internally, there can be incentives to distort verification (e.g., when seeking management promotions), resulting in cherry-picking charts with a perpetual upward march in skill. This is hardly unique to professionals in the weather forecasting enterprise, however. Sociology has widely studied the accountability

and transparency of institutions (Fung et al. 2007), offering strategies such as developing clear and measurable goals, providing regular and timely updates, using plain language, and soliciting feedback. However, currently, most operational weather services prioritize natural sciences over social sciences to a great extent.

Especially with the growing trend toward automation, evaluating meteorologists' contributions can be sensitive, creating tensions between operational meteorologists and the "forecasting police" (Pagano et al. 2016). Even Allen et al. (1952) mention the "certain meanness" in checking on individual forecasters (as opposed to having their tools, systems, training, and organizational context judged). Meteorologists want to make better forecasts and are motivated to test their operational beliefs/rules of thumb. Yet they can find themselves caught in the middle of ill-defined or misinterpreted forecast products, and narrowly defined key performance indicators (KPIs). In cases where downstream products are generated algorithmically (e.g., meteorologists issue grids which then generate text and icons), forecasters sometimes struggle to convey the true "weather story" consistently across the various products (Just and Foley 2020).

As forecast automation increases, the nature of operations-to-research feedback is evolving. Meteorologists are transitioning from making products to engaging customers in decision support (Stuart et al. 2022). However, customer-facing personnel can be distant from model developers (who may not even be in the same agency). Increased complexity or opacity of forecast-generating algorithms, especially with the rise of machine learning, makes it harder for humans to understand the reasons behind the forecast. When forecasters identify shortcomings in models, they often face uncertainty about whom to share this feedback with and how, if at all, the feedback will be used.

Some agencies face constraints that limit their ability to perform comprehensive verification. Smaller weather services, for example, often have limited staff, with only one or two meteorologists on duty, leaving little time for verification. Additionally, they lack data, bandwidth, appropriate software, and computing capabilities, further hindering evaluation efforts. Verification capacity can vary greatly, with large international centers conducting routine, novel, and complex evaluations using state-of-the-art software and different observations, presented in richly visual interfaces. Even resource-rich agencies experience challenges in maintaining systems, meeting tight operational deadlines, and handling the variable nature of weather.

Research opportunities. Can participatory exercises and data mining be used to build objective models of how experts subjectively assess performance (Crochemore et al. 2015)? Which verification approaches best educate individual forecasters about how to add value to objective guidance? Can machine learning and other approaches study the performance of meteorologists' forecast edits to model and improve these interventions? What performance measures should be publicly displayed and how should they be communicated? Which routinely reported KPIs provide the best actionable intelligence for management and policymakers? Can an operational real-time verification desk provide adequate feedback for improving forecast guidance? How can forecasting systems be holistically compared in a consistent way (e.g., an agency's NWP vs global centers vs third-party forecasters)? How can equitable verification services be provided to marginalized communities?

3. Calls to action

Ultimately, operational agencies seek to reap the benefits of verification activities by moving up the "Verification Information Value Ladder" (Fig. 5), transforming data into information and insights. They do this with the goal of impacting both forecaster and user decisions and

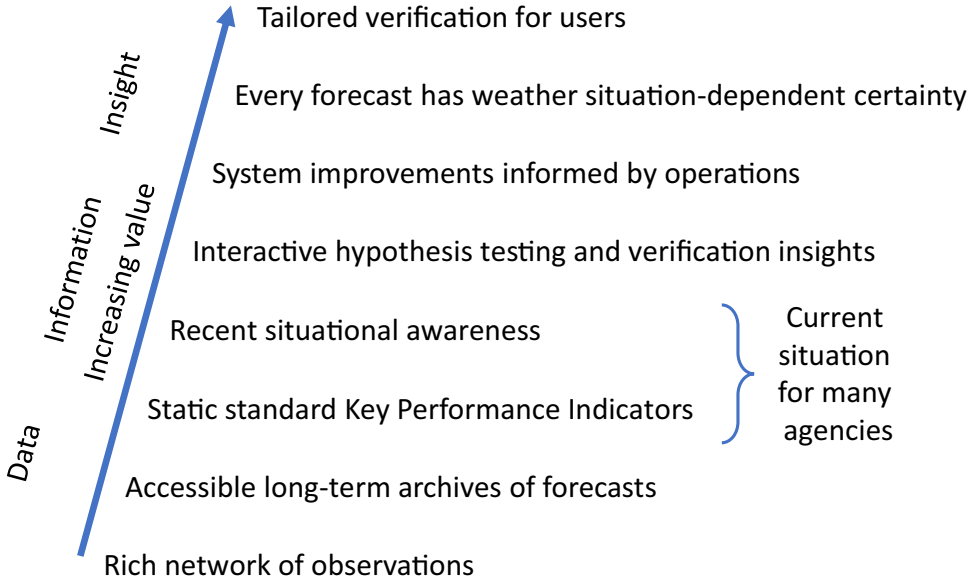

Fig. 5. Climbing the Verification Information Value Ladder: the foundations of verification systems are listed at the bottom, and increasingly valuable activities are listed near the top. The ultimate goal is to have well-integrated and accessible verification information that is tailored to user decisions. While certain numerical modeling centers have nearly achieved full maturity, many operational forecasting agencies have ample data and metrics but struggle to convert them into actionable insights.

developing appropriate levels of trust in the forecasts. In addition to the research opportunities listed above, we suggest paths forward to achieve this goal.

a. Define forecast products with verification in mind. Although more of an issue for agencies than academics, having an inadequately defined forecast product is the root of many verification challenges. For example, in Australia, wind forecasts historically provided either the average wind or the highest wind of the hour, depending on the context and the need to support fire danger products. The fire-weather focus negatively impacted wind power generators who sought consistent and unbiased products. The inconsistency also made it difficult to benchmark the official forecasts against automated guidance. The agency now produces three wind products ("on-hour," "max-in-hour," and "gust") to cater to diverse user needs. Although each of these products is single-valued, they are now defined within a probabilistic framework as the mean of the predictive distribution. Fundamentally, when forecast products are defined within a probabilistic framework like this, there is a solid theoretical basis (e.g., Gneiting 2011) for selecting appropriate verification metrics, making verification straightforward.

Probabilistic forecasts also have the added benefit of addressing the very strong user appetite for forecasts that state their certainty (Morss et al. 2008). Statements of forecast certainty greatly improve decision-making and lead to greater trust and more understanding of forecast information. These statements could take many forms, such as probabilistic forecasts, a range of scenarios, or presenting historical verification results next to the current forecasts. Although difficulties can arise when communicating probabilistic forecasts and their verification results, particularly to nonspecialist audiences, ample social science research demonstrates that users are adept at understanding risk concepts if framed in the right terms (National Academy of Sciences 2006). Social science can also help design forecast language and displays that minimize ambiguity and misinterpretations (Ripberger et al. 2022).

b. Enhance data availability. Currently, verification efforts face limitations due to inadequate observations, particularly high-quality measurements of impacts. To address this challenge,

countries should collaborate in creating low-latency comprehensive datasets that use standardized quality control, combine diverse sources of information (including carefully interpreted model analyses), provide complete coverage, and quantify certainty. This may involve data rescue, remote sensing, and partnerships with third parties, such as insurance companies. They should also work to improve the accessibility of data, such as by using common data formats and application programming interfaces (APIs).

In addition to improving observations, forecasting agencies should also prioritize the provision of hindcasts and historical as-issued forecasts to academics and users. By granting access to these datasets, individuals can conduct their own operationally relevant experiments. This collaborative approach empowers users to explore innovative methods and evaluate the performance of forecasts within their specific contexts, fostering a more comprehensive understanding of the strengths and weaknesses of different approaches.

c. Harness the power of shared software and systems. The benefits of community software for forecast verification are manifest: increased efficiency, standardization and compatibility, continuous improvement, and innovation. Already several operational agencies and academics have started to coalesce around certain systems such as the enhanced Model Evaluation Tools (METplus; Brown et al. 2021). METplus is a suite of open-source verification software developed by the National Center for Atmospheric Research Developmental Testbed Center. This modular and extensible Python-based framework offers a broad range of traditional verification metrics, ensemble verification, spatial verification, and other methods. The system has been used at operational agencies in the United States, United Kingdom, Australia, New Zealand, and South Africa and also universities, private companies, and nonprofits. It has enabled the intercomparison of spatial verification methods and serves as an effective way to transfer verification research into operations.

Additionally, the ability to share computer code is unprecedented, with myriad Python and R-based packages available to support weather forecast verification but also dataset handling, visualization, and others. Furthermore, there are many packages for metric calculation in other fields, such as hydrology, climate, statistics, or machine learning. As such packages proliferate, however, there is a need for better collaboration and consolidation to avoid diffuse redundant effort.

The rapid rise of cloud computing is also changing the nature of operational systems. Uploading data to cloud services is generally inexpensive, but downloading data is expensive. This makes verification an ideal cloud application given that the output datasets are often many orders of magnitude smaller than the input datasets. This may benefit countries with limited computing capacity, enable users or intermediaries to do their own verification, and make datasets more discoverable.

d. Foster an international verification community. We propose the development of an international collaborative Performance Assessment, Verification, and Evaluation (PAVE) community, inclusive of academics from social and natural science, operational agencies, forecast users, and the private sector. Operational personnel with limited resources and many demands on their attention rarely have the resources to read academic papers. Similarly, academics are under institutional pressures that can hinder their ability to do applications-oriented outreach. However, there is value to making resources available to help both communities traverse the "Valley of Death" of transitioning research to operations.

In addition to advocating for freely shared and jointly developed datasets and systems like those mentioned above, PAVE would support testbeds of focused effort organized around the operationally relevant challenges outlined in this paper (complementing the calls of Mass 2023). Special emphasis should be placed on reaping the benefits of the findings of social

science research and user engagement while also encouraging new research. Remarkably, there are hundreds of papers on communicating probability information but nearly none on communicating verification (Ripberger et al. 2022). The testbeds should have clear mechanisms for sharing constructive "operations to research" feedback to system and model developers. Cross-institutional engagement has increased in recent years, and international virtual conferences are becoming common, and so it is particularly important to include (and in some cases center) marginalized communities in collaborative research. PAVE could also play a significant role in verification training and outreach. For example, it would be valuable to enable those who understand scientific advances in the statistics and economics literature to write articles which makes those advances accessible to meteorological practitioners. Furthermore, those who have had success in using verification information could more widely share case studies and examples.

The Joint Working Group on Forecast Verification (JWGFVR, https://community.wmo.int/en/activity-areas/wwrp/wwrp-working-groups/wwrp-forecast-verification-research) has decades of valuable contributions, supporting the verification needs of the WMO and World Weather Research Programme's projects through experiments, workshops, and tutorials. There are evident synergies and opportunities for mutual support between PAVE, the JWGFVR, and other WMO projects. The complementary value of PAVE is in its outreach and engagement with broader audiences to address the operational challenges discussed herein. Those interested in collaborating with this community can contact the authors.

Acknowledgments. We are grateful for Lily Gao's bibliographic support and Melanie Erler's contributions to the article's figures. We also appreciate the constructive suggestions made by the anonymous peer reviewers as well as the Bureau of Meteorology's scientists Brendan Dimech, Andy Taylor, and Anja Schubert.

Data availability statement. No new data were created or analyzed in this study.

References

Allen, R. A., G. W. Brier, I. I. Gringorten, J. C. S. McKillip, C. P. Mook, G. P. Wadsworth, and W. G. Leight, 1952: Panel discussion on forecast verification (Held by the District of Columbia Branch on December 12, 1951). *Bull. Amer. Meteor. Soc.*, **33**, 274–278, https://doi.org/10.1175/1520-0477-33.7.274.

Bannister, T., and Coauthors, 2021: A pilot forecasting system for epidemic thunderstorm asthma in southeastern Australia. *Bull. Amer. Meteor. Soc.*, **102**, E399–E420, https://doi.org/10.1175/BAMS-D-19-0140.1.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Brown, B., and Coauthors, 2021: The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Amer. Meteor. Soc.*, **102**, E782–E807, https://doi.org/10.1175/BAMS-D-19-0093.1.

Casati, B., and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.*, **15**, 3–18, https://doi.org/10.1002/met.52.

——, M. Dorninger, C. A. S. Coelho, E. E. Ebert, C. Marsigli, M. P. Mittermaier, and E. Gilleland, 2022: The 2020 International Verification Methods Workshop online: Major outcomes and way forward. *Bull. Amer. Meteor. Soc.*, **103**, E899–E910, https://doi.org/10.1175/BAMS-D-21-0126.1.

Crochemore, L., C. Perrin, V. Andréassian, U. Ehret, S. P. Seibert, S. Grimaldi, H. Gupta, and J.-E. Paturel, 2015: Comparing expert judgement and numerical criteria for hydrograph evaluation. *Hydrol. Sci. J.*, **60**, 402–423, https://doi.org/10.1080/02626667.2014.903331.

Ebert, E., and Coauthors, 2013: Progress and challenges in forecast verification. *Meteor. Appl.*, **20**, 130–139, https://doi.org/10.1002/met.1392.

Fleischhut, N., S. M. Herzog, and R. Hertwig, 2020: Weather literacy in times of climate change. *Wea. Climate Soc.*, **12**, 435–452, https://doi.org/10.1175/WCAS-D-19-0043.1.

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

Fung, A., M. Graham, and D. Weil, 2007: *Full Disclosure.* Cambridge University Press, 304 pp.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Gneiting, T., 2011: Making and evaluating point forecasts. *J. Amer. Stat. Assoc.*, **106**, 746–762, https://doi.org/10.1198/jasa.2011.r10138.

——, and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831.

Golding, B., Ed., 2022: *Towards the "Perfect" Weather Warning.* Springer International Publishing, 270 pp.

Hartmann, H. C., T. C. Pagano, S. Sorooshian, and R. Bales, 2002: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.*, **83**, 683–698, https://doi.org/10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2.

Hogan Carr, R., K. Semmens, B. Montz, and K. Maxfield, 2021: Improving the use of hydrologic probabilistic and deterministic information in decision-making. *Bull. Amer. Meteor. Soc.*, **102**, E1878–E1896, https://doi.org/10.1175/BAMS-D-21-0019.1.

Joslyn, S., and S. Savelli, 2021: Visualizing uncertainty for non-expert end users: The challenge of the deterministic construal error. *Front. Comput. Sci.*, **2**, 590232, https://doi.org/10.3389/fcomp.2020.590232.

Just, A., and M. Foley, 2020: Streamlining the graphical forecast process. *J. South. Hemisphere Earth Syst. Sci.*, **70**, 108–113, https://doi.org/10.1071/ES19047.

Lenhardt, E. D., R. N. Cross, M. J. Krocak, J. T. Ripberger, S. R. Ernst, C. L. Silva, and H. C. Jenkins-Smith, 2020: How likely is that chance of thunderstorms? A study of how national weather service forecast offices use words of estimative probability and what they mean to the public. *J. Oper. Meteor.*, **8**, 64–78, https://doi.org/10.15191/nwajom.2020.0805.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: Extreme events and forecast evaluation. *Stat. Sci.*, **32**, 106–127, https://doi.org/10.1214/16-STS588.

Losee, J. E., K. Z. Naufel, L. Locker, and G. D. Webster, 2017: Weather warning uncertainty: High severity influences judgment bias. *Wea. Climate Soc.*, **9**, 441–454, https://doi.org/10.1175/WCAS-D-16-0071.1.

Lupo, K. M., C. S. Schwartz, and G. S. Romine, 2023: Displacement error characteristics of 500-hPa cutoff lows in operational GFS forecasts. *Wea. Forecasting*, **38**, 1849–1871, https://doi.org/10.1175/WAF-D-22-0224.1.

Marsigli, C., and Coauthors, 2021: Review article: Observations for high-impact weather and their use in verification. *Nat. Hazards Earth Syst. Sci.*, **21**, 1297–1312, https://doi.org/10.5194/nhess-21-1297-2021.

Mass, C., 2023: The uncoordinated Giant II: Why U.S. operational numerical weather prediction is still lagging and how to fix it. *Bull. Amer. Meteor. Soc.*, **104**, E851–E871, https://doi.org/10.1175/BAMS-D-22-0037.1.

Morss, R. E., J. L. Demuth, and J. K. Lazo, 2008: Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Wea. Forecasting*, **23**, 974–991, https://doi.org/10.1175/2008WAF2007088.1.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

——, S. Lichtenstein, B. Fischhoff, and R. L. Winkler, 1980: Misinterpretations of precipitation probability forecasts. *Bull. Amer. Meteor. Soc.*, **61**, 695–701, https://doi.org/10.1175/1520-0477(1980)061<0695:MOPPF>2.0.CO;2.

National Academy of Sciences, 2006: *Completing the Forecast.* National Academies Press, 112 pp.

Pagano, T. C., F. Pappenberger, A. W. Wood, M.-H. Ramos, A. Persson, and B. Anderson, 2016: Automation and human expertise in operational river forecasting. *Wiley Interdiscip. Rev.: Water*, **3**, 692–705, https://doi.org/10.1002/wat2.1163.

Palerme, C., M. Müller, and A. Melsom, 2019: An intercomparison of verification scores for evaluating the sea ice edge position in seasonal forecasts. *Geophys. Res. Lett.*, **46**, 4757–4763, https://doi.org/10.1029/2019GL082482.

Ripberger, J., A. Bell, A. Fox, A. Forney, W. Livingston, C. Gaddie, C. Silva, and H. Jenkins-Smith, 2022: Communicating probability information in weather forecasts: Findings and recommendations from a living systematic review of the research literature. *Wea. Climate Soc.*, **14**, 481–498, https://doi.org/10.1175/WCAS-D-21-0034.1.

Rodwell, M. J., J. Hammond, S. Thornton, and D. S. Richardson, 2020: User decisions, and how these could guide developments in probabilistic forecasting. *Quart. J. Roy. Meteor. Soc.*, **146**, 3266–3284, https://doi.org/10.1002/qj.3845.

Sherman-Morris, K., J. C. Senkbeil, and C. Vaughn, 2022: How close is close enough? A discussion of the distances relevant to personalizing tornado risk. *Bull. Amer. Meteor. Soc.*, **103**, E1573–E1586, https://doi.org/10.1175/BAMS-D-21-0142.1.

Stuart, N. A., and Coauthors, 2022: The evolving role of humans in weather prediction and communication. *Bull. Amer. Meteor. Soc.*, **103**, E1720–E1746, https://doi.org/10.1175/BAMS-D-20-0326.1.

Taggart, R., 2022: Evaluation of point forecasts for extreme events using consistent scoring functions. *Quart. J. Roy. Meteor. Soc.*, **148**, 306–320, https://doi.org/10.1002/qj.4206.

Turner, D. D., H. Cutler, M. Shields, R. Hill, B. Hartman, Y. Hu, T. Lu, and H. Jeon, 2022: Evaluating the economic impacts of improvements to the High-Resolution Rapid Refresh (HRRR) numerical weather prediction model. *Bull. Amer. Meteor. Soc.*, **103**, E198–E211, https://doi.org/10.1175/BAMS-D-20-0099.1

Wang, R. Y., and D. M. Strong, 1996: Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, **12**, 5–33, https://doi.org/10.1080/07421222.1996.11518099.

World Meteorological Organization, 2000: Guidelines on performance assessment of public weather services. WMO/TD-1023, 67 pp., https://library.wmo.int/idurl/4/44889.

——, 2015: Valuing weather and climate: Economic assessment of meteorological and hydrological services. WMO-1153, 308 pp., https://library.wmo.int/idurl/4/54637.