BAMS
Meeting Summary

The 2020 International Verification Methods Workshop Online

Major Outcomes and Way Forward

Barbara Casati, Manfred Dorninger, Caio A. S. Coelho, Elizabeth E. Ebert, Chiara Marsigli, Marion P. Mittermaier, and Eric Gilleland

ABSTRACT: The International Verification Methods Workshop was held online in November 2020 and included sessions on physical error characterization using process diagnostics and error tracking techniques; exploitation of data assimilation techniques in verification practices, e.g., to address representativeness issues and observation uncertainty; spatial verification methods and the Model Evaluation Tools, as unified reference verification software; and meta-verification and best practices for scores computation. The workshop reached out to diverse research communities working in the areas of high-impact weather, subseasonal to seasonal prediction, polar prediction, and sea ice and ocean prediction. This article summarizes the major outcomes of the workshop and outlines future strategic directions for verification research.

2020 International Verification Methods Workshop

What: The international forecast verification research community, together with diverse research communities working in the areas of high-impact weather, subseasonal to seasonal prediction, polar prediction, and sea ice and ocean prediction, met to share recent aspects of research on verification methods for environmental prediction and to discuss as well as further develop and promote best verification practices.

When: 9–13 and 16–20 November 2020

Where: Virtual, hosted by Barbara Casati (Environment and Climate Change Canada) jointly with Manfred Dorninger (University of Vienna).

KEYWORDS: Forecast verification/skill

https://doi.org/10.1175/BAMS-D-21-0126.1
Corresponding author: Barbara Casati, barbara.casati@ec.gc.ca
In final form 12 May 2021

AFFILIATIONS: Casati—Meteorological Research Division, Environment and Climate Change Canada, Dorval, Quebec, Canada; **Dorninger**—University of Vienna, Vienna, Austria; **Coelho**—Center for Weather Forecast and Climate Studies (CPTEC), National Institute for Space Research (INPE), Cachoeira Paulista, São Paulo, Brazil; **Ebert**—Bureau of Meteorology, Docklands, Victoria, Australia; **Marsigli**—Deutscher Wetterdienst, Offenbach am Main, Germany, and Arpae Emilia-Romagna, Bologna, Italy; **Mittermaier**—Met Office, Exeter, United Kingdom; **Gilleland**—Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

An International Verification Methods Workshop (IVMW) is held approximately every three years to present recent aspects of verification research and discuss best verification practices. The workshops are organized by the Joint Working Group on Forecast Verification Research (JWGFVR), which is joint between the World Meteorological Organization's (WMO) World Weather Research Programme (WWRP) and the Working Group for Numerical Experimentation (WGNE). The focus of these workshops is on the verification methodology, rather than on the application of methods to specific datasets. The broad remit of the workshops attracts participants from the operational environment, research institutions, and forecast user communities to discuss the application of novel verification methods to different types of weather forecasts and environmental predictions, spanning all spatial and temporal scales.

In November 2020 the JWGFVR held a virtual event, the "International Verification Methods Workshop Online" (2020-IVMW-O), in order to fill the gap between face-to-face workshops during the global lockdown caused by the COVID-19 pandemic. The workshop spanned two weeks and included 20 two-hour sessions, staggered across different time zones to accommodate presenters and attract participants from different countries around the globe. Session topics included physical error characterization using process diagnostics and error tracking techniques; exploitation of data assimilation (DA) techniques in verification; the Model Evaluation Tools (MET); spatial verification methods; and meta-verification and best practices for scores computation. Contributions from the WWRP High Impact Weather (HIWeather) project, Subseasonal to Seasonal prediction project (S2S), the Polar Prediction Project (PPP), and the sea ice and ocean (OceanPredict) communities were numerous, constituted by targeted sessions, and led to fruitful interdisciplinary exchanges. The program, abstract booklet, and presentations of the 2020-IVMW-O can be found at the workshop website (https://jwgfvr.univie.ac.at). The workshop closed with a discussion session which summarized its major outcomes, which are mirrored in this article. Where possible, reference is made to the relevant publications of the presenters.

Major scientific outcomes

Error tracking and process diagnostics. Error tracking, ensemble sensitivity analyses, and subsequent relaxation and/or observation system (data denial) experiments are dynamical approaches to the error characterization (see Magnusson 2017; Jung et al. 2014; Lawrence et al. 2019; Quinting and Grams 2022; Quinting et al. 2022; and presentations by S. Shields, M. Totha, Z. Wang). These techniques analyze the model error propagation in association with large-scale circulation, to identify regions and sources of major forecast busts. Results of these techniques are similar to those obtained by a verification conditioning on weather types, by using composites, applying principal component analysis or considering teleconnections. It

was discussed whether it might be possible to exploit artificial intelligence (AI) to automate these techniques, for example, for analyzing systematic errors and identifying forecast busts, or for clustering weather types for conditional verification on large-scale regimes. However, interpretation cannot be automated, and a strong educational component is still needed to correctly interpret error tracking and ensemble sensitivity analyses.

Process diagnostics focus on verifying the relationships between multiple variables, which mirrors the physical process(es) interrelating such physical variables. As an example, Baker et al. (2021) analyzed the land–atmosphere interactions, by selecting some diagnostics describing the processes relating soil moisture, soil temperature, evapotranspiration, and precipitation (based on previous literature), and then verifying the representation of these relationships spatially. Day et al. (2020) verified the snow–atmosphere interactions, by analyzing the response of surface and 2m air temperatures to radiative forcing (as in Miller et al. 2018), via scatterplots and regression lines. Similarly, A. Solomon presented an assessment of the representation of stable boundary layers by verifying the relationship between surface stability and the net long-wave radiation, again via scatterplots, conditioning both for clear sky or cloudy conditions. It was highlighted that most processes can be described by a simplified process-fitting function (e.g., a regression line or an index inter-relating multiple variables). Verification of these process-fitting functions and indices provides an assessment on how models represent the processes.

In the discussion it was acknowledged that process diagnostics are both multivariate and multidimensional, analyzing signals not only on two-dimensional fields, but also for vertical profiles and time series. Given the multivariate nature of process diagnostics, future development should investigate multivariate statistical tools. Collocated monitoring instruments for several process-relevant variables, beyond traditional meteorological variables (e.g., surface fluxes or radiation in the vertical column), provide an ideal observation dataset for process diagnostics. Process diagnostics can be improved by conditional verification, which can help stratify to account for different surface conditions or large circulation regimes (e.g., cloudy versus clear sky, mirroring low versus high pressure systems) affecting the processes.

The error tracking and process diagnostics sessions were the most attended of the conference. The JWGFVR aims to collaborate more closely with WGNE to further develop diagnostics which target physical processes and help identifying the sources of forecast errors.

Data assimilation, representativeness, and observation uncertainty. T. Auligné gave an overview of data assimilation (DA) techniques which can be exploited for verification purposes. The DA community has wide knowledge of observations and the associated uncertainties with tools to address representativeness, as well as established quality control (QC) procedures. There are also well-established tools for matching model output and observations, from simple interpolation to forward operators which simulate, e.g., radiances as measured by satellites. Moreover, DA routinely performs an analysis of the observation-model departures (for several observing systems) and the propagation in space and time of the increments, which can itself be an error source diagnostic (similar to error tracking). On the other hand, DA can benefit from spatial (field-morphing) verification techniques (Nehrkorn et al. 2014, 2015), as well as metrics for assessing model systematic errors and the impact of observation system experiments (OSEs). There is a desire from the JWGFVR and Data Assimilation and Observing Systems (DAOS) communities to enhance this two-way exchange.

Representativeness can dominate forecast errors, especially for coarse resolution models (which are limited by their resolution to represent subgrid-scale phenomena), and for highly discontinuous (e.g., precipitation, fog) or surface variables driven by local effects. North et al. (2022) showed how comparison of point versus collocated gridded observations provides a simple yet informative approach to analyze representativeness and

measure the effects on verification results. Ben Bouallegue (2020) and Ben Bouallegue et al. (2020) estimate the representativeness error by fitting parametric models to paired point and area-averaged observations; verification is then performed with a perturbed ensemble, where the estimated representativeness uncertainty is added to the ensemble uncertainty. B. Casati presented a lapse-rate adjustment of model temperatures to station elevations to address the altitude representativeness mismatching. In the discussion it was noted that some approaches to address representativeness issues can become as complex as postprocessing techniques. Nevertheless, data treatment is always necessary prior to verification, in order to correctly match model and observations; standardization of these preprocessing approaches was deemed not strictly necessary, as long as there is transparency in the processing that is applied. Finally, it was recognized that representativeness is a separate issue from observation uncertainty.

Observation uncertainty, its effects on verification results, and how to account for it in verification practices was the most recurrent topic of debate at the workshop. With the past decades of increasing improvements of NWP systems, the uncertainties inherent in the verifying observations are now comparable to the magnitude of forecast errors. Observation uncertainty can no longer be neglected in verification practices. Known measurement errors need to be addressed, since the benefits of the correction overcome the added uncertainty (Køltzow et al. 2020), while observation uncertainties from unknown sources need to be estimated and incorporated in verification statistics computation (Ferro 2017). There is a strong desire for a more complete understanding of the quality and error characteristics of the observing systems and verification datasets, not only for gridded fields, but also for time series and vertical profiles.

METplus: Toward a unified reference verification software. The capabilities of the Model Evaluation Tools (MET; Brown et al. 2021; www.dtcenter.org/community-code/model-evaluation-tools-met) and METplus (which connects MET, METviewer, METdatabd, and includes Python wrappers and embeddings), were showcased on diverse applications, including convection-allowing model development guidance and evaluation of air-quality forecasts. MET and METplus include many verification methods, from traditional deterministic and probabilistic scores against station (point) measurements, to more recent spatial verification methods against gridded observations, and tools for statistical inference, including confidence intervals. METplus enables customized input data and graphical display output; moreover, the GitHub platform enables new developments directly on the suite. Given its modularity and flexibility, open access, and community-based nature, METplus is becoming a repository of verification techniques, and can be used not only in research but also by operational weather services. In fact, METplus has been adopted as the operational verification software by NOAA/NCEP/NWS, as well as by the U.K. Met Office Unified Model Partnership.

An animated discussion revealed how much the verification community strives for universal reliable gridded reference observation datasets. Verification against (SYNOP) station networks is becoming cumbersome due to the station representativeness issues, and because station networks do not sample the verification domain homogeneously. Verification against gridded observations and/or analyses partially addresses these issues, and enables more sophisticated spatial diagnostics. However, the major drawback for verification against own-model analyses remains the background model dependence (Park et al. 2008). D. Hotta presented a twin analysis approach to quantify the effects of the error correlation (forecast-truth correlates with analysis-truth), when performing own-analysis verification. Verification against a multimodel ensemble of analyses was also discussed (Bowler et al. 2017a,b), in consideration also of the fact that single analyses still underestimate uncertainty (Bauer et al. 2014). Several two-dimensional observational datasets which are not model dependent

already exist (e.g., Contractor et al. 2020; Steinacker et al. 2006) which could be more fully exploited for verification. Accounting for the weaknesses and strengths of each observation dataset in the interpretation of the verification results is still fundamental.

Spatial verification methods. Spatial verification methods (Brown et al. 2012) continue to evolve, reflecting the very active community in this branch of verification research. The keynote by B. Brown and M. Dorninger gave an historical overview, including the early development of MET (first software hosting spatial techniques), and the spatial verification method intercomparisons (Gilleland et al. 2010; Dorninger et al. 2018). From a recent intercomparison of distance metrics (Gilleland et al. 2020) and better understanding of their shortcomings, Gilleland (2021) proposed new spatial alignment performance measures, and G. Skok presented the fraction skill score distance and distance neighborhood skill score. Neighborhood approaches are being further developed at MeteoFrance (Stein and Stoop 2019; Stein and Stoop 2021, manuscript submitted to *Mon. Wea. Rev.*), with the evaluation of within-neighborhood continuous ranked probability score (CRPS) and categorical scores which relax the location matching for the contingency table counts within neighborhoods. Novel object-based methods for verifying predictions of thunderstorm tracks (Skinner et al. 2018; Flora et al. 2019; Potvin et al. 2020) and extreme convective precipitation features (B. Sass, S. Anderson) were also presented. Finally, Brunet et al. (2012) illustrated Structure Similarity Indices and Metrics, which are widely used in image quality assessment (Wang and Bovik 2002, 2006; Wang et al. 2004). The verification community can benefit from adopting these scores, which merge scale-separation and distance metrics, summarize the performance in a single index, and allow error decomposition, while being robust and mirroring human judgement.

Spatial verification approaches are also exploited for process understanding. Borderies et al. (2018) expanded the concept of neighborhood approaches to vertical profiles by matching radar profiles to the most resembling modeled vertical profile within a radius of 160 km: this model-to-observed vertical profile matching enables separating and disentangling location errors from the process analysis. Griffin et al. (2020) applied the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a,b) to verify the NWP representation of clouds and water vapor by using all-sky infrared brightness temperatures from satellites. In both studies the verification was performed in observation space, by using a forward operator to simulate radar reflectivity and satellite brightness temperature, respectively.

Discussion ensued on the operationalization, aggregation, and inference of spatial verification methods. Most developments focus on case studies, whereas the techniques need to be adapted also for allowing aggregation on multiple cases and provide confidence intervals on the statistics. It was commented that, for gridded products, inference ought to account for the spatial interdependence between nearby grid points/locations. The advantage of having spatial verification approaches encoded in METplus is to provide intuitive automated spatial verification that can run systematically in NWP operational centers, with encoded standardized aggregating and inference tools. A lack of representation of novel spatial verification methods for ensemble prediction systems was noted. Future developments could address the spread–error relationship spatially (as in Dey et al. 2014), and spatial methods which address ensemble precipitation forecasts are needed, especially in operational context. Observation uncertainty remains to be addressed with spatial verification methods.

Meta-verification: Improved and new scores. The meta-verification sessions focused mainly on novel applications and refinement of existing statistics. Acharya and Tippet (2020) proposed the point-biserial correlation (Kornbrot 2014; Glass and Hopkins 1995) for verification of probabilistic forecasts for a naturally dichotomous phenomenon (e.g., precipitation

or clouds). Taggart (2020) introduced the Huber (1964) loss and Huber mean, a scoring rule which optimize the balance between quadratic penalties, for small errors, and linear penalties, for large errors. Bennett et al. (2021) and M. A. Sharpe proposed new methods for verifying the timing of events. North et al. (2022) showed the added information value of the stable equitable error in probability space (SEEPS; Rodwell et al. 2010; Haiden et al. 2012), by using satellite precipitation observations and scoring with an almost-global coverage. In the discussion SEEPS was recognized as a well-designed score for model development guidance, and it was recommended to explore the use of SEEPS for other variables (e.g., clouds, surface winds, high-impact variables such as extreme near-surface temperatures).

For ensembles, a wider use of the logarithmic ignorance score (Roulston and Smith 2002) and the generalized discrimination score (Mason and Weigel 2009; Weigel and Mason 2011) was recommended, in conjunction with the traditional Brier score, CRPS, reliability diagram and relative operating characteristic (ROC) curve. K. Nelson and Y. Ning proposed the use of information-theory based scores. In alignment with recent meta-verification literature, a few studies focused on a better understanding of the score properties and refining their decomposition and computation, e.g., with respect to binning (Leutbecher and Haiden 2020; Dimitriadis et al. 2021; Mitchell 2020; and presentation by H. E. Brooks). During the discussion, the practicality of providing R statistical packages to accompany these refinements in score computation was highlighted (e.g., scoringRules, Jordan et al. 2019). It was recognized that for research purposes R is statistically stronger, whereas Python is better suited for processing large data volumes (e.g., for operational verification). Finally, the verification community can benefit from the strengths of both languages (hence, no need to make an explicit choice) as the two languages can be interlaced.

Engaging diverse scientific communities

High-impact weather and the user value chain. The HIWeather project is exploring methods for evaluating the end-to-end warning value chain, linking observations and modeling to impacts, warnings, and community benefit (Zhang et al. 2019). The needs of the forecast user, whether it be operational meteorologists, emergency managers, industry, or the public, should be considered when designing a verification approach. For example, tailored verification of fire spread in coupled fire-atmosphere modeling provides useful information for fire forecasters and emergency managers in Colorado (presentation by A. Siems-Anderson). Consideration of users' perspectives in the design and evaluation of new services led to the successful development and implementation of new early warning system in Argentina (presentation by J. Chasco). B. G. Brown showed how model ranking, scorecards, and outlier examination provides more meaningful and useful information to hurricane center managers and forecasters concerning the relative performance of different cyclone models. To encourage greater use of ensembles by forecasters, Du et al. (2019) presented a composite score that combines ensemble mean error, spread, nonlinearity, and existence of outliers into a single number, and introduced the predictability horizon diagram index (PHDI), which measures an ensemble's temporal convergence toward the correct (observed) solution.

Sharpe et al. (2018) verify extremes where the thresholds identifying the distribution tails and skill are computed with respect to the local climatology. In the discussion it was noted that the method could be revisited by fitting theoretical distributions from extreme value theory (Coles 2001).

In an innovative twist, Rodwell et al. (2020) adapted the cost–loss model to represent user satisfaction, pain, regret, or thrill from acting (or not) on a forecast, where different users have their own probability thresholds at which they would make a decision. The expected value for a set of forecasts and integrated across users' decision thresholds can be formulated into a "user Brier score" that, unlike the traditional Brier score, does not assume a uniform

distribution of cost/loss ratios. A highlight of the workshop was an interactive experiment involving all the participants who, for a given probabilistic forecast (of strong winds), shared their decision thresholds for different scenarios ranging from sporting events on the beach, windmill operations, cycling, and exercising in a smoggy city.

Nontraditional observations are increasingly being used to evaluate forecasts of high-impact weather phenomena. For example, F. San Martino exploited crowdsourcing such as Twitter to verify gusts and hail events; D. Wilke made use of building damage reports to verify strong wind forecasts. Marsigli et al. (2021) described how the verified phenomena might be defined by different thresholds and/or related variables (e.g., forecast convective available potential energy and observed lightning flash density). Using multiple types of observations can characterize and reduce observational uncertainty, and fuzzy/spatial approaches can accommodate the non-exact spatial/temporal matching between forecasts and observations. Greater collaboration with the nowcasting community would be beneficial. The community seeks more automated approaches to filter and quality control reports from media, Twitter, insurance claims, impact assessments, emergency callouts, and crowdsourcing so they can be used to assess high-impact weather forecasts.

Subseasonal to seasonal and longer climate simulations and predictions. Subseasonal verification research has been promoted by the S2S prediction project (Coelho et al. 2018, 2019; de Andrade et al. 2019). F. Doblas-Reyes highlighted the importance of forecast quality for climate services, emphasizing the need for standard procedures and accounting for observations uncertainty. Manrique-Suñén et al. (2020) discussed verification challenges under subseasonal prediction systems heterogeneity. Sources of S2S precipitation forecast predictability were analyzed by using empirical orthogonal functions (presentation by C. A. S. Coelho), a regression approach (de Andrade et al. 2021), and teleconnection patterns (Lenssen et al. 2020). D. Büeler et al. (2021) discussed the role of flow-dependent verification and calibration for subseasonal weather regimes. N. Georgas presented climate products validation, highlighting the need for using emerging observing technologies. J. P. French proposed a permutation test for investigating distribution differences in climate simulations and reanalysis.

Discussions emphasized the need to use multiple observations or reanalyzes and account for associated uncertainties, for adequate verification interpretation. Analyses ensemble spread can have similar magnitude as forecast ensemble spread (Dorninger et al. 2018), highlighting the need for quantifying observational uncertainty. Scores comparing two ensemble distributions (from reanalyzes and models) are suitable for estimating and incorporating uncertainty in verification (e.g., Goessling and Jung 2018). Investigating spatial methods in climate verification is another avenue for exploitation.

Sea ice verification. Sea ice verification research has flourished in the context of the Polar Prediction Project, and several new spatial verification methods have been developed in recent years (Dukhovskoy et al. 2015; Melsom et al. 2019; Goessling et al. 2016; Goessling and Jung 2018; Mohammadi-Aragh et al. 2020; Linow and Dierking 2017). At the workshop A. Cheng presented a novel symmetric distance metric for the verification of ice-edge. Niraula and Goessling (2021) introduced a reference forecast which uses anomaly damped persistence as a benchmark for dynamical sea ice models. Model intercomparisons (Zampieri et al. 2018; Palerme et al. 2019) provided realistic testbeds for comparing the new sea ice verification approaches. Peterson et al. (2021, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*) showed how these new sea ice metrics are already being transferred to operational practices. Discussion extended to importing sea ice metrics into METplus. Future challenges in sea ice verification are the evaluation of sea ice thickness and pressure, with application to navigation safety.

Large discrepancies exist between satellite-based gridded observational datasets used for sea ice verification. Assessing the uncertainty of such remote-sensed datasets was recognized as an urgent task yet to be addressed.

Ocean verification. F. Hernandez and G. Smith showcased the activities of the intercomparison and validation Task Team of Ocean Predict (https://oceanpredict.org). The OceanPredict community has well-established reference verification datasets and basic verification statistics, which are exchanged routinely between operational centers for monitoring and comparing ocean analyses and forecasting systems. Y. Le Clainche showed an example of the assessment of the vertical profiles of water temperature and salinity for the Canadian ocean models. Similarly, E. Clementi showed operational verification of sea surface temperature and sea level anomalies for the Copernicus Marine Environment and Monitoring Services (CMEMS) ocean models over the Mediterranean and Black Sea basins.

The OceanPredict community is reaching out to the verification research community to advance the development of more sophisticated research-oriented approaches. Two pioneering spatial approaches were presented: Smith and Fortin (2022) introduced an object-based verification technique for ocean eddies in the Gulf Stream; Mittermaier et al. (2021) applied MODE to verify simulated chlorophyll-a blooms against a satellite ocean color product. Discussion also highlighted the strong interest in diagnostics for the ocean-atmosphere coupling processes (with and without sea ice), which would require synergies between ocean, atmosphere, cryosphere, and verification research communities.

Visualization tools. Web-based platforms and software-tools with visualization capability are essential for better understanding and communicating verification results. At the workshop examples were shown by A. Bentley for METplus; J. R. M. Garcia introduced the Model Evaluation Comparator; A. Paxian illustrated the DWD user-oriented climate prediction verification website (www.dwd.de/climatepredictions). Muñoz et al. (2019) presented an interface for IRI's Climate Predictability Tool (Mason et al. 2021).

Conclusions

The November 2020 International Verification Methods Workshop Online highlighted recent advancements in verification research and the way forward. There is strong momentum to exploit error tracking techniques and further develop, in concert with the modeling community, diagnostics which target physical processes and help to identify the sources of weather and climate prediction errors. Similarly, there is the desire to enhance synergies with the data assimilation community, e.g., to address representativeness issues and quantify the observation uncertainties, and incorporate those in the computation of verification scores. MET/METplus has the capability of becoming the unified reference verification software, including state-of-the-art verification and inference statistics, suitable both for research and operational environments.

The Spatial Verification session outlined the uptake of these methods in operational environments as well as for research diagnostic tools, and the link with structure similarity indices and metrics was highlighted. Future avenues of research could expand spatial verification methods to ensembles and account for observation uncertainty in spatial approaches. The operational use of SEEPS against gridded observations and beyond precipitation was also promoted. In the meta-verification sessions, the point-biserial correlation, Huber loss and logarithmic score were reproposed, for verification practices.

User-oriented verification and the evaluation of the end-to-end value chain, from observations and modeling to impacts and warnings, was illustrated for several science-to-services applications. The use of nontraditional observations, including crowdsourcing, is also being

exploited for high-impact weather verification. Seasonal to decadal climate prediction verification practice has been consolidating in the past decade through standard procedures (World Meteorological Organization 2010), which will soon include subseasonal standards. In the context of PPP, the sea ice community has been developing several spatial methods assessing different sea ice attributes. The ocean community is striving for new verification approaches, beyond the already well-established baseline score exchange, such as diagnostics to assess the atmosphere-ocean coupling.

The attendance of this virtual event was more than double that of past face-to-face IVMW, and the quality of the abstracts and presentations was outstanding. Moreover, because of the practical ease of online participation (without travel time commitments and expenses), the outreach to diverse research communities was very large. Given this remarkable outcome, the JWGFVR has decided to alternate face-to-face and online IVMW on a biannual basis, even post-pandemic.

Acknowledgments. The authors wish to acknowledge Pete Saddler and his team at Environment and Climate Change Canada for the 2020-IVMW-O technical support with MS Teams, and Markus Ristic at University of Vienna for setting up the workshop website.

References

Acharya, N., and M. K. Tippet, 2020: Point-biserial correlation-based skill scores for probabilistic forecasts. Earth and Space Science Open Archive, https://doi.org/10.1002/essoar.10505449.1.

Baker, J. C. A., and Coauthors, 2021: An assessment of land–atmosphere interactions over South America using satellites, reanalysis and two global climate models. *J. Hydrometeor.*, **22**, 905–922, https://doi.org/10.1175/JHM-D-20-0132.1.

Bauer, P., L. Magnusson, J.-N. Thepaut, and T. M. Hamill, 2014: Aspects of ECMWF model performances in polar areas. *Quart. J. Roy. Meteor. Soc.*, **142**, 583–596, https://doi.org/10.1002/qj.2449.

Ben Bouallegue, Z., 2020: Accounting for representativeness in the verification of ensemble forecasts. ECMWF Tech. Memo. 865, 21 pp., www.ecmwf.int/node/19544.

——, T. Haiden, N. J. Weber, T. M. Hamill, and D. S. Richardson, 2020: Accounting for representativeness in the verification of ensemble precipitation forecasts. *Mon. Wea. Rev.*, **148**, 2049–2062, https://doi.org/10.1175/MWR-D-19-0323.1.

Bennett, J. C., D. E. Robertson, Q. J. Wang, M. Li, and J.-M. Perraud, 2021: Propagating reliable estimates of hydrological forecast uncertainty to many lead times. *J. Hydrol.*, **603**, 126798, https://doi.org/10.1016/j.jhydrol.2021.126798.

Borderies, M., and Coauthors, 2018: Simulation of W-band radar reflectivity for model validation and data assimilation. *Quart. J. Roy. Meteor. Soc.*, **144**, 391–403, https://doi.org/10.1002/qj.3210.

Bowler, N. E., and Coauthors, 2017a: Inflation and localization tests in the development of an ensemble of 4D-ensemble variational assimilations. *Quart. J. Roy. Meteor. Soc.*, **143**, 1280–1302, https://doi.org/10.1002/qj.3004.

——, and Coauthors, 2017b: The effect of improved ensemble covariances on hybrid variational data assimilation. *Quart. J. Roy. Meteor. Soc.*, **143**, 785–797, https://doi.org/10.1002/qj.2964.

Brown, B. G., E. Gilleland, and E. E. Ebert, 2012: Forecasts of spatial fields. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed., I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 274 pp.

Brown, B. G., 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.

Brunet, D., E. R. Vrscay, and Z. Wang, 2012: On the mathematical properties of the structural similarity index. *IEEE Signal Process. Lett.*, **21**, 1488–1499, https://doi.org/10.1109/TIP.2011.2173206.

Büeler, D., L. Ferranti, L. Magnusson, L. F. Quinting, and C. M. Grams, 2021: Year-round sub-seasonal forecast skill for Atlantic–European weather regimes. *Quart. J. Roy. Meteor. Soc.*, **147**, 4283–4309, https://doi.org/10.1002/qj.4178.

Coelho, C. A. S., M. A. Firpo, and F. M. de Andrade, 2018: A verification framework for South American sub-seasonal precipitation predictions. *Meteor. Z.*, **27**, 503–520, https://doi.org/10.1127/metz/2018/0898.

——, B. G. Brown, L. Wilson, M. P. Mittermaier, and B. Casati, 2019: Forecast verification for S2S time scales. *Sub-Seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*, 1st ed. A. W. Robertson and F. Vitart, Eds., Elsevier, 585 pp.

Coles, S., 2001: *An Introduction to Statistical Modeling of Extreme Values.* Springer, 208 pp.

Contractor, S., and Coauthors, 2020: Rainfall Estimates on a Gridded Network (REGEN): A global land-based gridded dataset of daily precipitation from 1950 to 2016. *Hydrol. Earth Syst. Sci.*, **24**, 919–943, https://doi.org/10.5194/hess-24-919-2020.

Davis, C. A., B. G. Brown, and R. G. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

——, —— and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, https://doi.org/10.1175/MWR3146.1.

Day, J. J., and Coauthors, 2020: Measuring the impact of a new snow model using surface energy budget process relationships. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002144, https://doi.org/10.1029/2020MS002144.

de Andrade, F. M., C. A. S. Coelho, and I. F. Cavalcanti, 2019: Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dyn.*, **52**, 5451–5475, https://doi.org/10.1007/s00382-018-4457-z.

——, M. P. Young, D. MacLeod, L. C. Hirons, S. J. Woolnough, and E. Black, 2021: Sub-seasonal precipitation prediction for Africa: Forecast evaluation and sources of predictability. *Wea. Forecasting*, **36**, 265–284, https://doi.org/10.1175/WAF-D-20-0054.1.

Dey, S. R. A., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev.*, **142**, 4091–4107, https://doi.org/10.1175/MWR-D-14-00172.1.

Dimitriadis, T., T. Gneiting, and A. Jordan, 2021: Stable reliability diagrams for probabilistic classifiers. *Proc. Natl. Acad. Sci. USA*, **118**, e2016191118, https://doi.org/10.1073/pnas.2016191118.

Dorninger, M., E. Gilleland, B. Casati, M. P. Mittermaier, E. E. Ebert, B. G. Brown, and L. Wilson, 2018: The setup of the MesoVICT project. *Bull. Amer. Meteor. Soc.*, **99**, 1887–1906, https://doi.org/10.1175/BAMS-D-17-0164.1.

Du, J., B. Zhou, and J. Levit, 2019: Measure of forecast challenge and predictability horizon diagram index for ensemble models. *Wea. Forecasting*, **34**, 603–615, https://doi.org/10.1175/WAF-D-18-0114.1.

Dukhovskoy, D. S., J. Ubnoske, E. Blanchard-Wrigglesworth, H. R. Hiester, and A. Proshutinsky, 2015: Skill metrics for evaluation and comparison of sea ice models. *J. Geophys. Res. Oceans*, **120**, 5910–5931, https://doi.org/10.1002/2015JC010989.

Ferro, C. A. T., 2017: Measuring forecast performance in the presence of observation error. *Quart. J. Roy. Meteor. Soc.*, **143**, 2665–2676, https://doi.org/10.1002/qj.3115.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental warn-on-forecast system. *Wea. Forecasting*, **34**, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

Gilleland, E., 2021: Novel measures for summarizing high-resolution forecast performance. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **7**, 13–34, https://doi.org/10.5194/ascmo-7-13-2021.

——, D. A. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365–1373, https://doi.org/10.1175/2010BAMS2819.1.

——, G. Skok, B. G. Brown, B. Casati, M. Dorninger, M. P. Mittermaier, N. Roberts, and L. J. Wilson, 2020: A novel set of verification test fields with application to distance measures. *Mon. Wea. Rev.*, **148**, 1653–1673, https://doi.org/10.1175/MWR-D-19-0256.1.

Glass, G. V., and K. D. Hopkins, 1995*: Statistical Methods in Education and Psychology.* 3rd ed. Allyn & Bacon, 674 pp.

Goessling, H. F., and T. Jung, 2018: A probabilistic verification score for contours: Methodology and application to Arctic ice-edge forecast. *Quart. J. Roy. Meteor. Soc.*, **144**, 735–743, https://doi.org/10.1002/qj.3242.

——, S. Tietsche, J. J. Day, E. Hawkins, and T. Jung, 2016: Predictability of the Arctic sea ice edge. *Geophys. Res. Lett.*, **43**, 1642–1650, https://doi.org/10.1002/2015GL067232.

Griffin, S. M., J. A. Otkin, G. Thompson, M. Frediani, J. Berner, and F. Kong, 2020: Assessing the impact of stochastic perturbations in cloud microphysics using *GOES-16* infrared brightness temperatures. *Mon. Wea. Rev.*, **148**, 3111–3137, https://doi.org/10.1175/MWR-D-20-0078.1.

Haiden, T., M. J. Rodwell, D. S. Richardson, A. Okagaki, T. Robinson, and T. D. Hewson, 2012: Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Mon. Wea. Rev.*, **140**, 2720–2733, https://doi.org/10.1175/MWR-D-11-00301.1.

Huber, P., 1964: Robust estimation of a location parameter. *Ann. Math. Stat.*, **35**, 73–101, https://doi.org/10.1214/aoms/1177703732.

Jordan, A., F. Kruger, and S. Lerch, 2019: Evaluating probabilistic forecasts with scoringRules. *J. Stat. Software*, **90**, 1–37, https://doi.org/10.18637/jss.v090.i12.

Jung, T., M. A. Kasper, T. Semmler, and S. Serrar, 2014: Arctic influence on subseasonal midlatitude prediction. *Geophys. Res. Lett.*, **41**, 3676–3680, https://doi.org/10.1002/2014GL059961.

Køltzow, M., B. Casati, T. Haiden, and T. Valkonen, 2020: Verification of solid precipitation forecasts from numerical weather prediction models in Norway. *Wea. Forecasting*, **35**, 2279–2292, https://doi.org/10.1175/WAF-D-20-0060.1.

Kornbrot, D., 2014: Point biserial correlation. *Wiley StatsRef: Statistics Reference Online*, N. Balakrishnan et al., Eds., Wiley, https://doi.org/10.1002/9781118445112.stat06227.

Lawrence, H., N. Bormann, I. Sandu, J. J. Day, J. Farnan, and P. Bauer, 2019: User and impact of Arctic observations in the ECMWF Numerical Weather Prediction system. *Quart. J. Roy. Meteor. Soc.*, **145**, 3432–3454, https://doi.org/10.1002/qj.3628.

Lenssen, N., L. Goddard, and S. Mason, 2020: Seasonal forecast skill of ENSO teleconnection maps. *Wea. Forecasting*, **35**, 2387–2406, https://doi.org/10.1175/WAF-D-19-0235.1.

Leutbecher, M., and T. Haiden, 2020: Understanding changes of the continuous ranked probability score using a homogeneous Gaussian approximation. *Quart. J. Roy. Meteor. Soc.*, **147**, 425–442, https://doi.org/10.1002/qj.3926.

Linow, S., and W. Dierking, 2017: Object-based detection of linear kinematic features in sea ice. *Remote Sens.*, **9**, 493, https://doi.org/10.3390/rs9050493.

Magnusson, L., 2017: Diagnostic methods for understanding the origin of forecast errors. *Quart. J. Roy. Meteor. Soc.*, **143**, 2129–2142, https://doi.org/10.1002/qj.3072.

Manrique-Suñén, A., N. Gonzalez-Reviriego, V. Torralba, N. Cortesi, and F. J. Doblas-Reyes, 2020: Choices in the verification of S2S forecasts and their implications for climate services. *Mon. Wea. Rev.*, **148**, 3995–4008, https://doi.org/10.1175/MWR-D-20-0067.1.

Marsigli, C., and Coauthors, 2021: 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.

Mason, S. J., and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **137**, 331–349, https://doi.org/10.1175/2008MWR2553.1.

——, M. K. Tippet, L. Song, and Á. G. Muñoz, 2021: Climate Predictability Tool version 17.3.2. Columbia University Academic Commons, https://doi.org/10.7916/d8-pxg1-yr77.

Melsom, A., C. Palerme, and M. Müller, 2019: Validation metrics for ice edge position forecasts. *Ocean Sci.*, **15**, 615–630, https://doi.org/10.5194/os-15-615-2019.

Miller, N. B., M. D. Shupe, J. T. M. Lenaerts, J. E. Kay, G. de Boer, and R. Bennartz, 2018: Process-based model evaluation using surface energy budget observations in Central Greenland. *J. Geophys. Res. Atmos.*, **123**, 4777–4796, https://doi.org/10.1029/2017JD027377.

Mitchell, K. 2020: Score decompositions in forecast verification. Ph.D. thesis, University of Exeter, 163 pp., http://hdl.handle.net/10871/40923.

Mittermaier, M. P., R. North, J. Maksymczuk, C. Pequignet, and D. Ford, 2021: Using feature-based verification methods to explore the spatial and temporal characteristics of the 2019 chlorophyll-*a* bloom season in a model of the European Northwest Shelf. *Ocean Sci.*, **17**, 1527–1543, https://doi.org/10.5194/os-17-1527-2021.

Mohammadi-Aragh, M., M. Losch, and H. F. Goessling, 2020: Comparing Arctic Sea ice model simulations to satellite observations by multiscale directional analysis of linear kinematic features. *Mon. Wea. Rev.*, **148**, 3287–3303, https://doi.org/10.1175/MWR-D-19-0359.1.

Muñoz, Á. G., and Coauthors, 2019: PyCPT: A Python interface and enhancement for IRI's Climate Predictability Tool. Zenodo, https://zenodo.org/badge/latestdoi/142679292.

Nehrkorn, T., B. Woods, T. Auligné, and R. N. Hoffman, 2014: Application of feature calibration and alignment to high- resolution analysis: Examples using observations sensitive to cloud and water vapor. *Mon. Wea. Rev.*, **142**, 686–702, https://doi.org/10.1175/MWR-D-13-00164.1.

——, B. Woods, R. N. Hoffman, and T. Auligné, 2015: Correcting for position errors in variational data assimilation. *Mon. Wea. Rev.*, **143**, 1368–1381, https://doi.org/10.1175/MWR-D-14-00127.1.

Niraula, B., and H. F. Goessling, 2021: Spatial damped anomaly persistence of sea-ice edge for reference forecast. *J. Geophys. Res. Oceans*, **126**, e2021JC017784. https://doi.org/10.1029/2021JC017784.

North, R. C, M. P. Mittermaier and S. F. Milton, 2022: Using SEEPS with a TRMM-derived climatology to assess global NWP precipitation forecast skill. *Mon. Wea. Rev.*, **150**, 135–155, https://doi.org/10.1175/MWR-D-20-0347.1.

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

Park, Y.-Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.*, **134**, 2029–2050, https://doi.org/10.1002/qj.334.

Potvin, C. K., and Coauthors, 2020: Assessing systematic impacts of PBL schemes in the NOAA warn-on-forecast system. *Mon. Wea. Rev.*, **148**, 2567–2590, https://doi.org/10.1175/MWR-D-19-0389.1.

Quinting, J. F., and C. M., Grams, 2022: EuLerian Identification of ascending AirStreams (ELIAS 2.0) in numerical weather prediction and climate models – Part 1: Development of deep learning model. *Geosci. Model Dev.*, **15**, 715–730, https://doi.org/10.5194/gmd-15-715-2022.

——, ——, A. Oertel, and M., Pickl, 2022: EuLerian Identification of ascending AirStreams (ELIAS 2.0) in numerical weather prediction and climate models – Part 2: Model application to different datasets. *Geosci. Model Dev*, **15**, 731–744, https://doi.org/10.5194/gmd-15-731-2022.

Rodwell, M. J., D. S. Richardson, T. D. Hewson, and T. Haiden, 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **136**, 1344–1363, https://doi.org/10.1002/qj.656.

——, 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.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660, https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Sharpe, M. A., C. E. Bysouth, and R. L. Stretton, 2018: How well do Met Office post-processed site-specific probabilistic forecasts predict relative-extreme events? *Meteor. Appl.*, **25**, 23–32, https://doi.org/10.1002/met.1665.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype warn-on-forecast system. *Wea. Forecasting,* **33,** 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Smith, G. C., and A. -S. Fortin, 2022: Verification of eddy-properties in operational oceanographic analysis systems. *Ocean Modell.*, in press.

Stein, J., and F. Stoop, 2019: Neighborhood-based contingency tables including errors compensation. *Mon. Wea. Rev.*, **147**, 329–344, https://doi.org/10.1175/MWR-D-17-0288.1.

Steinacker, R., and Coauthors, 2006: A mesoscale data analysis and downscaling method over complex terrain. *Mon. Wea. Rev.*, **134**, 2758–2771, https://doi.org/10.1175/MWR3196.1.

Taggart, R., 2020: Point forecasting and forecast evaluation with generalised Huber loss. Bureau Research Rep. 050, Bureau of Meteorology, 29 pp., www.bom.gov.au/research/publications/researchreports/BRR-050.pdf.

Wang, Z., and A. C. Bovik, 2002: A universal image quality index. *IEEE Signal Process. Lett.*, **9**, 81–84, https://doi.org/10.1109/97.995823.

——, and ——, 2006: *Modern Image Quality Assessment.* Morgan & Claypool Publishers, 156 pp.

——, ——, H. R. Sheikh, and E. P. Simoncelli, 2004: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, **13**, 600–612, https://doi.org/10.1109/TIP.2003.819861.

Weigel, A. P., and S. J. Mason, 2011: The generalized discrimination score for ensemble forecasts. *Mon. Wea. Rev.*, **139**, 3069–3074, https://doi.org/10.1175/MWR-D-10-05069.1.

World Meteorological Organization, 2010: Manual on the global data-processing and forecasting systems. WMO-485, 123 pp., https://library.wmo.int/doc_num.php?explnum_id=10164.

Zampieri, L., H. F. Goessling, and T. Jung, 2018: Bright prospects for Arctic Sea ice prediction on subseasonal time scales. *Geophys. Res. Letters*, **45**, 9731–9738, https://doi.org/10.1029/2018GL079394.

Zhang, Q., and Coauthors, 2019: Increasing the value of weather-related warnings. *Sci. Bull.*, **64**, 647–649, https://doi.org/10.1016/j.scib.2019.04.003.