SCORECARDS

for Convection-Allowing Models

Since scientists first began modeling the Earth system, a need to verify the subsequent forecasts has existed. Formal verification can identify ways forecast models need to improve, objectively judge how changes in the models affect forecast quality, and provide the best set of metrics to evaluate for different users. While improvements in numerical weather prediction are measured by objective statistics, the selection of appropriate statistics necessarily incorporates subjectivity. A set of agreed-upon verification metrics would aid in assessing forecast quality and in determining the impact of changes to numerical models.

Operational implementation of new guidance occurs only after a thorough evaluation of the strengths and weaknesses of forecasts compared to observations and previous model iterations. Given the large complexity and dimensionality of most atmospheric forecast problems, care must be taken when selecting the statistics to verify.

Global models with resolutions on the scale of tens of kilometers use scores that summarize synoptic aspects of the forecast such as placement and intensity of high- and low-pressure systems. Convection-allowing models (CAMs), with typical grid spacing of ~3 km, simulate finescale features that need not necessarily exactly collocate with the observed features to be valuable, and so verification metrics must allow spatial

and/or temporal displacement to determine the full value of the forecast.

The difficulty with assessing a multitude of statistical scores is that often optimizing one score will degrade another. Alternately, improving a score for one model field may reduce the same statistic in another field. For example, tests of the upgrade of the Global Forecast System (GFS) to the Finite-Volume Cubed GFS (FV3GFS) showed that the upgrade improved the northward quantitative precipitation forecast bias in the GFS, but worsened the low bias in instability and 2-m dewpoint fields. Moreover, improvements may occur solely at certain forecast hours, requiring metrics from multiple times. Nuances and trade-offs that necessarily occur during model implementation may be inadvertently overlooked in this myriad of verification metrics, despite being relevant to one or more communities within the weather enterprise. A summary visualization tool—a scorecard—can display this large quantity of information to model developers and the meteorological community as a whole.

Scorecards compare model systems at multiple field thresholds, statistics, time periods, and domains in one image. Systemic differences can then be examined in depth in order to diagnose model deficiencies. For instance, if a new system

Adapted from “Initial Development and Testing of a Convection-Allowing Model Scorecard,” by **Burkely T. Gallo** (CIMMS/University of Oklahoma, and NOAA/NWS/NCEP Storm Prediction Center), **Christina P. Kalb**, **John Halley Gotway**, **Henry H. Fisher**, **Brett Roberts**, **Israel L. Jirak**, **Adam J. Clark**, **Curtis Alexander**, and **Tara L. Jensen**. Published online in *BAMS*, December 2019. For the full, citable article, see [DOI:10.1175/BAMS-D-18-0218.1](https://doi.org/10.1175/BAMS-D-18-0218.1).

Fields and stratifications deemed important by DTC workshop participants for CAM scorecards in general, for all applications.

Forecast field	Deterministic stratifications
Ceiling (column)	Forecast Length (0–36 h), Threshold (500, 1,000, 3,000, 5,000, 10,000 ft), Domain (west and east CONUS)
Visibility (surface)	Forecast Length (0–36 h), Threshold (0.5, 1.0, 3.0, 5.0, 10.0 mi), Domain (west and east CONUS)
Dewpoint (2 m)	Forecast Length (0–36 h), Diurnal (0000–2300 UTC), Domain (west and east CONUS)
Specific humidity (column)	Forecast Length (0–36 h), Diurnal (0000, 1200 UTC)
Temperature (2 m)	Forecast Length (0–36 h), Diurnal (0000–2300 UTC), Domain (west and east CONUS)
Temperature (column)	Forecast Length (0–36 h), Diurnal (0000, 1200 UTC)
Wind (10 m)	Forecast Length (0–36 h), Diurnal (0000–2300 UTC), Domain (west and east CONUS)
Wind (column)	Forecast Length (0–36 h), Diurnal (0000, 1200 UTC)
Precipitation (surface)	Forecast Length (0–36 h), Threshold (0.01, 0.1, 0.25, 0.5, 1.0, 2.0, 3.0, 4.0, 6.0 in., percentiles), Scale (3, 40 km), Domain (west and east CONUS)
Simulated reflectivity (composite)	Forecast Length (0–36 h), Threshold (25, 30, 35, 40, 45, 50 dBZ, percentiles), Scale (3, 40 km), Domain (west and east CONUS)
Updraft helicity (2–5 km; 0–3 km AGL)	Forecast Length (0–36 h), Threshold (25, 50, 75 m ² s ⁻² , percentiles), Scale (3, 40 km), Domain (W and E CONUS)

has difficulty with nocturnal temperatures, that would become evident from the evaluation at different times as displayed on the scorecard rather than potentially obscured by a summary metric evaluated over the entire forecast run. While the subjectivity of metric selection remains, careful selection of fields, metrics, and domains of most value to key end users can optimize the scorecard to form an overall picture of model performance.

Scorecards have previously compared upgrades to operational systems such as the Global Deterministic and Ensemble Forecast System, the impact of new data assimilation schemes, and aerosol impacts at the subseasonal time frame. These studies show the flexibility of the scorecard framework: different scorecards can be used for deterministic and ensemble forecasts and also can encompass metrics that concern different forecast interests.

Extending the scorecard framework to determine the best operational implementations of CAMs requires consideration and planning, as found during the initial development and implementation of a CAM scorecard for the 2018 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (SFE). The models that would be evaluated using the CAM scorecard were determined prior to the 2018 SFE. After the SFE (on 30 April–1 June), the process of determining appropriate fields, thresholds, and metrics took time and focused on problems of interest in the experiment. As with any verification study, a clear scientific problem should drive what the scorecard displays. The SFE focused on severe convection and required metrics indicating how well the model is forecasting heavy precipitation, high reflectivity cores, and rotating storms, with the later addition of variables that determine favorable storm environments. Other applications will likely need to display different fields on their scorecards, and a broad community engagement can ensure that scorecards for future operational CAM implementation include relevant fields and metrics for a variety of users.

METViewer CAM Scorecard

for NSSLFV3 and HRRR_nsslgrid

2018-04-30 00:00:00 - 2018-06-01 00:00:00

A scorecard comparing the bias and CSI of the temperature and 10-m *U* and *V* wind components between the NSSL-FV3 and the HRRRv3. Green (purple) colors indicate that the NSSL-FV3 scored higher (lower) than the HRRRv3.

			Daily Domain					CONUS				
			12 hr	18 hr	24 hr	30 hr	36 hr	12 hr	18 hr	24 hr	30 hr	36 hr
CSI	Temperature	>=32		△	▲	△		▽	▽	▽	▽	▽
		>=65	▽	▽	▽	▽	▽	▽	▽	▽	▽	▽
		>=70	▽	▽	▽	▽	▽	▽	▽	▽	▽	▽
		>=75	▽	▽	▽	▽	▽	▽	▽	▽	▽	▽
		>=80		▽	▽	▽		▽	▽	▽	▽	▽
	U Wind	>=10 kts			▽		▽	▽	▽	▽		
		>=15 kts					▽	▽				
		>=20 kts	▽	△		▲	▽	▲	△	△		
		>=25 kts	△	△		△	△		△		▲	
	V Wind	>=10 kts	▽					▽	▽	▽	▽	
		>=15 kts	▽		▽							
		>=20 kts			▽	▽		▽				
		>=25 kts	△	▽	▽		△	△				△
Bias	Temperature	>=32	▲	△		△		△		▽	△	△
		>=65	▽		△	▽	▽	▽	△	▽	▽	▽
		>=70	▽	△	△	▽	▽	▽	△	▽	▽	▽
		>=75	▽	▽	▽	▽	▽	▽	▽	▽	▽	▽
		>=80		▽	▽	▽		▽	▽	▽	▽	▽
	U Wind	>=10 kts	▲		▲			△	△	▲	▽	
		>=15 kts						▲		▽	▽	△
		>=20 kts					▽	△	△	▽	△	△
		>=25 kts	△				△			△	△	
	V Wind	>=10 kts				▽		▽	▽		▽	▽
		>=15 kts			▲			△	△		▽	▽
		>=20 kts						▲			▲	
		>=25 kts	△	▽								

△	NSSLFV3 is better than HRRR_nsslgrid at the 99% significance level
▲	NSSLFV3 is better than HRRR_nsslgrid at the 95% significance level
	No statistically significant difference between NSSLFV3 and HRRR_nsslgrid
▽	NSSLFV3 is worse than HRRR_nsslgrid at the 95% significance level
▽	NSSLFV3 is worse than HRRR_nsslgrid at the 99% significance level
	Not statistically relevant

CAM VERIFICATION NEEDS

To address CAM verification needs across the meteorological enterprise, two community-based working groups (the CAM and Verification and Validation Groups, established by NOAA) have combined their efforts. They are assisting NOAA with developing a strategic implementation plan for CAM verification as the United States transitions to a Unified Forecast System designed around the FV3 dynamical core. Developing unified metrics and verification strategies will enable critical evaluation of the Next-Generation Global Prediction System. With the working groups' recommendations, systemic and relevant evaluation will support an evidence-based advance of the future Unified Forecast System.

To determine the most important metrics and fields for evaluating CAMs, the two working groups created a spreadsheet of 30 forecast fields. Crucial details of the simulated fields such as vertical and temporal attributes, validation sources, potential stratifications, and needed statistical scores for both a deterministic and ensemble framework were considered. Participants at a summer 2018 Developmental Testbed Center (DTC) metrics workshop¹ assigned each model field a rating of priority (higher for fields with multiple applications) and readiness (availability of quality observations).

Only a certain number of rows in a scorecard can be displayed simultaneously. A scorecard with too many becomes unwieldy. Many subjective choices remain, such as what to evaluate, how to evaluate it, and at what thresholds. By combining priority and readiness, the initial 30 fields were winnowed to 11 initial fields with applications ranging from aviation to air quality to winter weather.

The workshop showed that myriad metrics and fields are important to the meteorological community. Developing a comprehensive scorecard that addresses all concerns for all applications may be impossible. As such, the workshop attendees also highlighted the need for multiple stakeholders to contribute to the selection of metrics and raised the possibility of different scorecards for different applications.

SCORECARD DEVELOPMENT

The scorecard itself is generated using the Model Evaluation Tools (MET), a suite of statistical tools that form a unified verification framework. MET was initially developed to replicate the Environmental

² METplus includes METviewer, an R statistics package-based visualization tool that generates scorecards—available to the community via GitHub

¹ The Developmental Testbed Center Community Unified Forecast System Test Plan and Metrics Workshop, 30 July–1 August 2018 in Silver Spring, MD, with participants from NOAA, NASA, the U.S. Navy and Air Force, private sector, universities, and international collaborators.

Modeling Center mesoscale verification system, and computes more than 85 different traditional statistics with confidence intervals. The flexible framework offers object-based verification metrics, complementing traditional, gridpoint-based metrics and providing a potential future direction for the CAM scorecard. MET is at the core of METplus,² a unified verification and diagnostic capability being developed for the Unified Forecast System. METplus streamlines generation of the CAM scorecard from a variety of model and observational data sources. The CAM scorecard, as with its counterparts for convection-parameterizing models, allows users to select scores relevant to their interests. This flexibility necessitates the ongoing discussion (begun at the DTC metrics workshop) about which metrics should be included. For real-time use, we recommend reserving a sufficient development period to ensure timely data availability for scorecard generation.

Due to the complexity of setting up the real-time scorecard, a very small subset of fields was chosen for the initial scorecard, with expansion planned for the 2019 SFE. Initial fields focused on severe weather forecasting: simulated reflectivity, accumulated precipitation, and a probabilistic surrogate severe field based on updraft helicity. For the selected fields, the scorecard visualizes multiple verification thresholds and for two domains: the full CONUS and a movable daily domain (8.72° latitude × 15° longitude) centered on where the most severe convective weather was expected.

During the 2018 SFE, the scorecard was presented during the Friday morning forecast discussion, allowing participants to match their subjective impressions of the models formed throughout the week with the objective verification provided by the scorecard. Participant impressions of the scorecard were generally favorable.

THE FUTURE OF THE CAM SCORECARD

After the 2018 SFE, planned upgrades to METplus include the addition of surrogate severe weather and percentile capabilities so that METplus can incorporate preprocessing of these data and eliminate steps that users currently have to complete. Working with the datasets generated during SFE 2018, statistics for additional environmental fields such as 2-m temperature and 10-m *U* and *V* wind components were included in the scorecard. These fields are of critical interest to severe convective

≡ METADATA

BAMS: What would you like readers to learn from this article?

Burkely T. Gallo: *I would like readers to learn how complex a relatively simple-seeming process like forecast verification can be. Hopefully this work provokes thought on how best to design verification studies and emphasizes that there likely is no “one-size-fits-all” approach to verification. What metrics and fields are evaluated must depend on the questions being asked.*

BAMS: What surprised you in making the scorecard?

BTG: *I was surprised by how much the color and symbols that we used influenced the interpretability of the scorecard. Feedback from participants in the Spring Forecast Experiment and reviewers of the manuscript really helped to improve the scorecard design.*

BAMS: What was the biggest challenge in the scorecard project?

BTG: *There were so many fields and metrics put forth at the DTC workshop. Despite our relatively narrow focus on severe weather, we still had difficulties figuring out exactly what to include on the scorecard without making it have too many rows!*

BAMS: What’s next? How will you follow up?

BTG: *During the 2019 Spring Forecast Experiment, we implemented the real-time scorecard for the second year. We will hopefully have the scorecard in the 2020 SFE as well, and may apply it to more models than we have in previous years. Also, the practically perfect and surrogate severe methods of generating smoothed fields above a threshold will be available in an upcoming release of MET.*

forecasters as well as the wider meteorological community. The use of categorical statistics for the 2-m temperature and winds demonstrates the utility of using scores beyond traditional continuous measures. For example, it demonstrates that during the 2018 SFE, HRRRv3 tended to perform better in cold temperatures within the domain, but NSSL-FV3 tended to have higher skill at warmer temperatures. Additionally, it appears that the NSSL-FV3 performed better at lower wind speed thresholds and HRRRv3 at higher ones.

These mixed results can be commonplace if enough different fields, metrics, and times are evaluated. It is exceptionally challenging to develop a new implementation of a model that exceeds the performance of the prior model in all ways. This is especially true when looking beyond severe weather. For example, the NWS is concerned about forecast problems ranging from air quality to winter weather to tropical systems. A scorecard for any one of these applications—let alone the entire enterprise—could have a plethora of rows and mixed results. It is therefore imperative to consider practical significance as well as statistical significance when comparing two modeling systems. The scorecard will rarely provide a clear “correct answer” across all aspects being evaluated.

The expansion of the CAM scorecard for the Spring Forecasting Experiments into environmental information is the initial effort toward having the scorecard encompass other meteorological scales and processes, and demonstrates how the CAM scorecard can distinguish between models that may be quite similar in aggregate statistics or for a smaller selection of metrics. The expansion process is expected to involve multiple stakeholders, as was done in the DTC Metrics Workshop. Combining perspectives from groups throughout the meteorological community can provide consistent judgement of new model implementations from upgrade to upgrade, and interested parties can hone in on metrics, fields, and thresholds important to them. By strengthening and fostering these partnerships during development, input from across the weather enterprise can be incorporated and the scorecard can be developed to best serve the community. Hopefully, the scorecard can provide a visualization tool for a unified framework that includes aspects of model performance important to both model developers and end users such as operational forecasters. ••

AMS CENTENNIAL MONOGRAPH

A tribute to the past century of innovation within our community and a source of inspiration for the scientists and researchers still to come.

A Century of Progress in Atmospheric and Related Sciences: Celebrating the American Meteorological Society Centennial marks 100 years of scientific research in the areas covered by AMS publications. The monograph will consist of 27 articles, which together review 100 years of progress in key fundamental areas of research and the grand challenges in those areas of research in the coming decades.

PRE-ORDER TODAY!

AMETSOC.ORG/100MONOGRAPH

The Thinking Person's Guide to Climate Change

Robert Henson

Expanded and updated from Henson's *Rough Guide to Climate Change*, 3rd edition (no longer in print), combining years of data with recent research, including conclusions from the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, the Guide breaks down the issues into straightforward categories:

- Symptoms, including melting ice and extreme weather
- Science, laying out what we know and how we know it
- Debates, tackling the controversy and politics
- Solutions and Actions for creating the best possible future

© 2014, 516 pages, paperback

ISBN: 978-1-878220-73-7

List price: \$30 AMS Member price: \$20

AMS BOOKS

RESEARCH APPLICATIONS HISTORY

➤ bookstore.ametsoc.org