tomorrow.io

Best Practices for Weather Forecast Validation: Executive Summary

March 2021

1. Summary

Forecast verification, or "validation", is the process of evaluating the quality of a numerical weather prediction (NWP and other types of weather prediction models e.g. NowCast) tool against real measurements, usually via analysis of statistics that characterize the relationship between forecast and observed weather variables of interest [*1*]. Depending on the attributes of the chosen dataset(s) in question, the data analyst will likely employ different sets of statistical approaches, which are detailed herein. Undertaking statistical validation studies thus enables one to determine the inherent value of a given NWP tool; when equipped with this intelligence, commercial businesses can make more-informed decisions toward accomplishing strategic objectives.

2. Background

Standard validation approaches have been established by within the global scientific community (e.g., the World Meteorological Organization, the U.S. National Oceanographic and Atmospheric Administration (NOAA), and the U.S. National Aeronautics and Space Administration) to promote consistency in forecast verification. Up front, it matters whether the target weather variable(s) typically take on "yes"/ "no" types of conditions (as in the case of precipitation occurrence–a "discrete" variable situation), as opposed to representing an infinite range of real numbers (e.g., near-surface temperature–a "continuous" variable situation). For both cases, there are several pieces of critical background for the data analyst to take into account before carrying out statistical validation.

Gaps or interruptions to routine weather observations are common in the field of atmospheric data science, therefore quality control procedures are vital during the preliminary analysis and data-ingest production phases (prior to producing a weather forecast).

As alluded to above, comparisons between numerical weather forecasts and observations (i.e., validation) can, and often should, involve an independent dataset to serve as a benchmark or statistical baseline. Moreover, since data from NWP systems and observation platforms may have different underlying resolution, initial data pre-processing can rectify differences in underlying units (e.g., to unintentional offsets between data samples). It may also be useful to enact data thresholding (e.g., to account for differing sensitivity of the chosen sample datasets), for example, by conditioning precipitation data on accumulation > 1 mm prior to analysis. These preliminary steps are essential precursors to validation.

3. Validation Approaches

3.1 Discrete Variables

For discrete variables, contingency tables (**Fig. 1**) have proven to be highly effective analysis frameworks, as forecast output can be readily evaluated against observations for accuracy [*2*]. The analysis proceeds by summing up the number of times that the forecast and model agree (Hits and Correct Rejections), as well as when the forecast and observations do not agree (False Alarms and Misses).

		Observed	
		Yes	*No*
Forecasted	*Yes*	Hits	False Alarms
	No	Misses	Correct Rejections

Figure 1. A traditional contingency table used for tabulating forecast-observation outcomes.

From these summaries, key performance indicators are Probability of Detection (POD) and False Alarm Ratios (FAR). The Critical Success

Index (CSI) characterizes how often the forecast accurately identifies target event occurrence. As an extension, the Brier Skill Score (BSS) vets the model forecast and observation against a third independent source of ground truth, i.e., climatology. Then to handle issues with spatial or temporal mismatch between model forecast and observation output, the Fractions Skill Score (FSS) can be used (*see Sec. 5 below*).

3.2 Continuous Variables

For continuous variable validation, there exists a relevant set of statistics that a data analyst uses to potentially

reveal more detailed aspects of a forecast model's skill [*3*]. These statistical metrics are chosen depending on the need to quantify either bias, the magnitude of the error between forecast and observation data, or the correlation (i.e., how well the forecasts track observations). Another set of statistical performance metrics are appropriate when the sample data distributions do not resemble the standard "Bell Curve" [*4*].

4. Case Study

Tomorrow.io strongly believes in its mission to collect better data, create better models, and provide a superior set of forecast products. As part of that mission, considerable time and effort is invested in routine verification for quality assurance, ensuring that the company's product suite meets technical specifications and maintains its position at the forefront of weather technology.

Here, we present the results from a set of real-world validation experiments that illustrate several of the techniques that have been described herein. The sample data for this study were aggregated over four months in 2019, at hourly intervals before being normalized to the same unit of measurement.

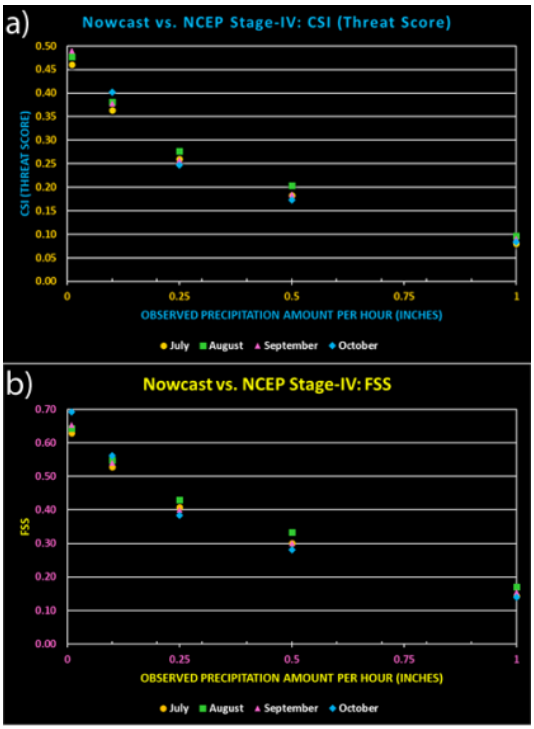

Figure 2. NowCast (NC) hour-1 forecast (a) CSI and (b) FSS over the continental United States for five one-hour precipitation accumulation thresholds (≥ 0.01 in. (0.25 mm), ≥ 0.10 in. (2.54 mm), ≥ 0.25 in. (6.35 mm), ≥ 0.50 in. (12.7 mm), and ≥ 1.0 in. (25.4 mm)). The NC data are validated using Stage-IV precipitation data from the National Centers for Environmental Prediction. Performance scores are shown for July (orange), August (green), September (pink), and October (blue) 2019.

Five operational thresholds of Tomorrow.io's NowCast precipitation accumulation were compared to NOAA's flagship precipitation verification dataset; **Figure 2** above shows that CSI and FSS metrics are well within the range of acceptance, as defined by NOAA's Weather Prediction Center.

5. Advanced Topics

When validation involves comparatively high-resolution NWP output paired with observations, more-advanced validation approaches are sometimes warranted to mitigate potential "double penalty" issues [*5*]. The "double penalty" results when the forecast is penalized two times for correctly capturing a weather feature of interest, but displacing it relative to

observations. "Neighborhood methods" set a boundary around the point-of-interest and seek to quantify the

fraction of local grid points that meet a specified event threshold.

Then using the FSS metric provides concise accounting for the difference in neighborhood probability of occurrence for the event of interest. The question of whether a forecast model exhibits significant skill is reduced to evaluating how close the FSS comes to a value of 1.0–a perfect forecast.

6. Closing

Since it is very difficult to treat the various sources of uncertainty associated with human-only observations (e.g., people often only consider a limited spatial area "outside the window" or "at a given location on the property"), it is critical to report weather conditions or parameters based on actual instrument measurement prior to validation. Analysts conventionally use direct or indirect meteorological measurements (e.g., real-time precipitation vs. sky cover), which are then formatted properly for subsequent analysis according to community-approved discrete and continuous variable validation methods described above. With the information contained in this brief, stakeholders will be better equipped to execute fair, objective validation and thereby realize the quality and value of Tomorrow.io's forecast products for their commercial operations and business decision-making purposes.

References

[1] Wilks, D.S., 2016: Statistical Methods in the Atmospheric Sciences (4th Ed.). Elsevier, 818 pp.

[2] *Forecast Verification*. Joint Working Group For Verification Research of the World Meteorological Organization (WMO) – World Weather Research Program, 2015, https://community.wmo.int/jwgfvr-verification. Accessed 27 February 2021.

[3] Brown, B., R. Bullock, T. Fowler, H. Gotway, J. K. Newman, T. Jensen, 2020: The MET Version 9.1.1 User's Guide. Developmental Testbed Center. Available at: https://github.com/dtcenter/MET/releases

[4] Stanski, H. R., L. J. Wilson. and W. R. Burrows, *Survey of Common Verification Methods in Meteorology (WMO World Weather Watch Technical Report No.8, WMO/TD No. 358) (2nd Ed.),* July 1989. Atmospheric Environment Service, Environment Canada, https://www.cawcr.gov.au/projects/verification/Stanski_et_al/Stanski_et_al.html. Accessed 27 February 2021.

[5] Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E., 2009: Intercomparison of Spatial Forecast Verification Methods, *Wea. and Forecast.*, 24(5), 1416-1430. Retrieved Mar 5, 2021, from https://journals.ametsoc.org/view/journals/wefo/24/5/2009waf2222269