tomorrow.io

Best Practices for Weather Forecast Validation: Technical Approach Guidelines for Data Analysts

March 2021

Table of Contents

I. Executive Summary

The purpose of this document is to describe technically accepted approaches to validate weather model forecasts. Validation is the process of evaluating the quality of a weather forecast, usually via analysis of statistics that characterize the relationship between forecast and observed weather variables of interest (*Wilks*, 2016). Validation may involve comparing output from one weather model to output from a second weather model, as well as comparing output from a weather model to real-world, *in-situ* observations.

For each type of validation above, there are several pieces of critical background for the data analyst to take into account before undertaking statistical evaluations. Gaps or interruptions to routine observation data recording are common in the field of atmospheric data science, therefore quality control procedures are vital during the preliminary analysis phase; enacting quality-control routines during the data ingest phase of production helps improve the accuracy of weather forecast initialization. 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. 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. Thresholding potentially reduces the data to encompass only a finite range of categorical data values, whereas an almost infinite range of forecast and observation data values are possible without any thresholding.

Given that weather variables can take on discrete values (in the former case just above) or continuous values (in the latter case), different methods are required in order to carry out validation. Contingency style approaches have been shown to be highly effective for quantifying the accuracy of models for which the output can be readily encoded/categorized (*WMO*, 2015). For continuous variable validation, there exists a relevant set of scalars and statistical indices which the data analyst conventionally uses in conjunction with sample probability distributions to determine the approximate accuracy/skill of a numerical weather forecast product. When comparatively high-resolution numerical weather prediction products are being vetted, issues stemming from spatial/temporal mismatch between sample datasets will often obfuscate the results of validation experiments; to mitigate these issues, some advanced validation techniques, such as neighborhood methods (e.g., *Roberts and Lean*, 2008; *Wilks*, 2016) may be necessary.

To illustrate the discussion points herein, two case studies are provided which compare Tomorrow.io's NowCast and Bespoke Atmospheric Model output to independent data sources as part of separate, exemplary validation exercises. Finally, comment on human-observation-only vs. direct/indirect observation as it pertains to validation is provided for data analysts to consider ahead of weather forecast validation.

1. Introduction

Tomorrow.io's weather forecasting and analysis products include a Current Conditions Layer (CCL) analysis, the NowCast (NC) short-term weather forecast, and the Tomorrow.io Bespoke Atmospheric Model that provides medium-range weather prediction. CCL is a data analysis system that assimilates many thousands of observations in real time and displays a fused layer that depicts the current state of the weather across the globe. The NC model produces highly detailed, short-term precipitation forecasts (otherwise known as precipitation "nowcasts"). Since its inception in 2017, several upgrades to NC have been deployed and the model has run continuously both in the United States and around the world, fed by a combination of real-time precipitation data from radars, satellites, and proprietary data sources where available. CBAM is a state-of-the-art Numerical Weather Prediction (NWP) system that provides custom short- and medium-range forecasts anywhere in the world. Tomorrow.io's science team and lead technologists regularly consult the available peer-review literature and attend technical conferences to maintain the company's cutting-edge technologies in-line with the accepted standards set forth by the global scientific community. We strongly believe in our mission to collect better data, create better models, and provide a superior set of forecast products.

As part of that mission, Tomorrow.io invests considerable time and effort in routine verification, ensuring that its product suite meets technical specifications, and toward carrying out validation for quality assurance. Forecast verification (i.e., validation) is the process of evaluating the quality of a forecast, usually via analysis of statistics that characterize the relationship between forecasted and observed weather variables of interest (Wilks, 2016). While the quality of a forecast can be evaluated based qualitatively on cursory review of attributes like bias, the generally accepted practice in academia and by the global atmospheric science community (e.g., according to the World Meteorological Organization, or WMO) is to conduct quantitative assessments of forecast accuracy and skill against a reference or baseline dataset. Yet, the exact approach to assess skill, and thereby validate a particular forecast product, depends on the type of prediction rendered.

The purpose of this document is to describe technically accepted approaches to validate forecasts of both *discrete* (where the forecasts take on one value out of a small, finite set of possible values at each output time/location) and *continuous* (where the forecasts can assume any number of real values within an extensive range of possibilities) weather variables. We will detail validation methodologies as well as briefly cover related assumptions, where necessary. We include real case studies on Tomorrow.io's forecast model performance relative to traditional reference datasets (e.g., produced by the National Oceanographic and Atmospheric Administration) as examples. With the information contained in this document, 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.

2. Validation Methodology

2.1. Preliminary Background

2.1.1. Quality Control (QC)

In data sparse regions or in areas of the world which present logistical challenges, it may be difficult for operators to maintain continuous meteorological observations in a consistent manner. This can result in limited data coverage, lacking period of record, as well as intermittent data gaps. For this reason, observation quality control (QC) strategies often incorporate gap-filling (e.g., using the local mean, sample, median, or other representative metric) and imputation (e.g., inferring values from proximate observations) to intelligently complete observation samples where missing. Thus, QC is a vital process for numerical weather prediction systems that ingest various observations before calculating their initial conditions (e.g., as part of the data assimilation), ahead of producing a forecast.

2.1.2. Remedying Data Inconsistency A Priori

As a complement to initial quality control procedures and before conducting validation studies, there may be situations where certain types of inconsistencies exist between the datasets of interest. It is therefore important to confirm that the units of each dataset match prior to comparison, otherwise artificial biases may become apparent in the results. The simplest case involves using constant conversion factors for magnitude/time/distance, as appropriate. With other incompatibilities, as with precipitation for example, analysts commonly work with metrics of total accumulation over a pre-defined time span (e.g., over an hour or a day) as opposed to working with instantaneous rates. For the latter, it is important to take the sampling interval into account when making any necessary conversions to adjust rate to total accumulation.

When comparing model output to point-data sources, it is also possible that data inconsistencies will exist in space and/or time. Care should be taken to ensure that the timestamps reference either Local Time (LT) or Greenwich Mean Time (GMT/UTC) together. Furthermore, point observations should be referenced as near as possible to the standard output times for numerical model output. This could imply sub-setting data from an individual surface observation station to obtain reports every three hours, every hour, on the top of the hour, or more frequently (depending on the temporal resolution of the datasets in question). Once temporal mismatch is accounted for, spatial inconsistencies can be addressed.

According to standard documentation for the National Center for Atmospheric Research Model Evaluation Tools (NCAR MET, e.g., *Brown et al.*, 2020), there are several ways to rectify spatial inconsistency between datasets. For differences between the specified height of point data and model output data above Earth's surface, linear interpolation between vertical grid points in altitude or linear interpolation using the natural logarithm of pressure is appropriate. When there are differences in the specified data locations in the horizontal directions, nearest neighbor/area-weighting/distance-weighting schemes are convenient methods to co-locate point data and model output data. For comparisons of grid-based datasets, re-gridding techniques can be used to pair data, effectively coarsening high-resolution data to match the positions of low-resolution data. For more detailed treatment of these concepts, the reader is referred to Chapter 5 (Re-Gridding), Chapter 7 (Point-based Statistical Tools), and Chapter 8 (Grid-based Statistical Tools) in *Brown et al.* (2020).

2.1.3. Establishing Baseline for Validation

A core tenet of validation practice is establishing a baseline for statistical comparison. This means that if two separate datasets are to be compared, it is best to seek out at least one additional independent source of quality-controlled data to facilitate the evaluation process (e.g., akin to objective third-party verification). The aforementioned reference to "independent" implies that the baseline dataset is not previously utilized by the data source(s) being validated. There are third-party, private vendors who provide quality-controlled data curation services for this purpose. In addition, research-grade meteorological data products are available from federal agencies like the National Oceanographic and Atmospheric Administration (NOAA), the National Aeronautics and Space Administration (NASA), or from federally funded research laboratories. A reliable source of "ground-truth" data are well-sited ground observation stations, such as those in the NOAA's automated weather observation system (AWOS) or surface observation station (ASOS) networks. One drawback however, is that AWOS/ASOS data are generally hosted only at commercial, regional, and sometimes at local airports. Furthermore, these stations are occasionally subject to local biases, such as channeling effects from winds interacting with nearby topography or being inadvertently situated close to sources of heat/moisture (e.g., biological, artificial, or otherwise).
Sometimes the best data for baseline comparisons comes from numerical modeling systems or reanalysis systems that use local/in-situ data augmentation approaches to improve the data product prior to dissemination. Again, this

discussion underscores the necessity of carrying out thorough QC/quality-assurance procedures and routine data "sanity checks" to ensure data reliability, as part of validation.

2.1.4. Thresholding

Thresholding–limiting data values to select ranges/intervals above or below some pre-determined level or rate–is one possible way toward conducting "conditional validation". This strategy may be required for a number of reasons including: to mitigate the impact of differences in nominal sensitivity of the observation(s)/modeling system(s) in question; to target a select, extreme class of rare events for analysis; or to enable dichotomous/binary classification (i.e., to designate either affirmative or negative occurrence for an event/phenomenon). As it will be detailed below, whether or not the analyst opts to utilize data thresholding can determine whether approaches for discrete or continuous validation would be appropriate to consider.

2.2 Validation Approach – Discrete Variables

There are instances where the determination of favorable vs. unfavorable business operation conditions depends on the choice between a single (or a small number) of "discrete" meteorological criteria. Therefore, it may be necessary to establish a categorical threshold. For example, real-world scenarios in the aviation sector include overcast vs. clear sky conditions or horizontal visibility greater than six statute miles. Precipitation observables can also be considered as "discrete", depending on how they are analyzed in practice:

- precipitationOccurrence [1, 0]
- precipitationAccumulation (e.g., based on thresholding relative to some minimum) [mm, in.]
- precipitationProbability [%]
- precipitationType [0: N/A, 1: Rain, 2: Snow, 3: Freezing Rain, 4: Ice Pellets]

For discrete variable validation, detection accuracy is a critical evaluation factor, as in, "did precipitation occur?" (yes/no); a proper categorical threshold is often based on empirical results or real-world classifications. Following from above, a "yes" forecast can be defined as an instance where the predicted precipitation rate exceeds 0.01 in. hr^{-1}, whereas a "no" forecast can be defined as times where no precipitation is forecasted (with similar conventions for the observation and benchmark datasets). These types of forecast-measurement data pairs can be recorded at individual point locations or within model grid-box areas.

2.2.1 The Contingency Table

According to the Joint Working Group For Verification Research (JWGFVR) of the World Meteorological Organization's (WMO) World Weather Research Program, a useful framework to quantify detection accuracy is the contingency table approach illustrated in **Figure 1** below (*WMO*, 2015). Note, contingency table assessments can be conditional on the presence or occurrence of some specific weather phenomenon of interest (e.g., convective vs. non-convective rainfall) (*Stanski et al.*, 1989).

Qualitatively speaking, an accurate forecast product will provide only predictions of Hits and Correct Negatives (i.e., correct identifications made for when the event does not occur) for a variable of interest. In contrast, a forecast product that produces Misses and False Alarms for some fraction (or all) of the time indicates that the particular guidance tool does not have a great deal of accuracy. In order to determine the accuracy of a forecast dataset in the context of discrete variable analysis, it is necessary to synthesize the information contained in the color-shaded regions of **Figure 1** below.

	Observed

		Yes	*No*	*Total*
Forecasted	*Yes*	Hits, n_{11}	False Alarms, n_{10}	# Yes Forecasts
	No	Misses, n_{01}	Correct Rejections, n_{00}	# No Forecasts
Total		# Yes Observations	# No Observations	$T = n_{00} + n_{01} + n_{10} + n_{11}$

Figure 1. A standard contingency table for classifying discrete meteorological variable outcomes (adapted from *WMO*, 2015).

2.2.2. Validation Statistics for Discrete Variables

There are a few key statistical measures to estimate detection accuracy according to the breakout of events captured in the contingency table (e.g., reference separate table elements in **Fig. 1**):

- Probability of Detection (POD) – What fraction of "yes" forecasts were correctly forecast? A "1" is a prefect forecast.

$$POD = \frac{n_{11}}{n_{01} + n_{11}}$$

- False Alarm Ratio (FAR) – What fraction of predicted "yes" forecasts did not actually occur? A "0" is a perfect forecast.

$$FAR = \frac{n_{10}}{n_{10} + n_{11}}$$

- Critical Success Index (CSI) – How well did the forecast "yes" events correspond to the observed "yes" events (i.e., ignoring the null case where no event was forecast) overall? A "1" is a perfect forecast.

$$CSI = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

- Frequency Bias (FB) – How did the frequency of forecast "yes" events compare to the observed frequency of "yes" events? Closer to 1 is better and the bias can be less than 1 (under-forecast) or greater than 1 (over-forecast).

$$FB = \frac{n_{11} + n_{10}}{n_{11} + n_{01}}$$

- Accuracy (ACC) – What fraction of the forecasts were correctly identified as Hits or Correct Rejections? Closer to 1 is better.

$$ACC = \frac{n_{11} + n_{00}}{T}$$

Note for rare events, true accuracy may be inflated because the forecast can simply revert to never predicting the event of interest and yet still obtain a high ACC score. It is best to pair consideration of ACC with other performance metrics defined above, or the combination of metrics that results in defining the Receiver Operating Characteristic (ROC) curve. The ROC curve illustrates the series of ratios of POD to FAR (note, both of these statistics range between 0 and 1) for various user-defined event thresholds (e.g., precipitation accumulation or precipitation rate thresholds,

etc.). Forecast skill improves as the area-under-the-ROC-curve (i.e., AUC) approaches 1.0, whereas values close to 0.0 indicate poor forecast performance. Following the JWGFVR, an of AUC = 0.5 indicates that the forecasted number of Hits equals the number of False Alarms overall, or essentially that the forecast has no skill (*WMO*, 2015).

An extension of discrete variable analysis considers the probability that events will or will not occur. For validating products based on probabilistic predictions, the Brier Score (i.e., BS; *Bradley et al.*, 2008) is particularly useful. By definition, this metric accumulates errors in probability predictions made by forecast and climatology. Then the Brier Skill Score can be used to assess forecast skill against climatology or the chosen baseline dataset. Statistical analysis software modules with built-in functions to compute this aforementioned collection of discrete variable statistics are available from within the NCAR MET as well as in common scientific software packages (e.g., from Python and R).

2.3. Validation Approach – Continuous Variables

In contrast to dealing with discrete variables and categorical thresholding, when the range of possible values for a target variable of interest becomes large, a different approach is warranted to appropriately quantify forecast model skill (*Wilks*, 2016). For instance, accuracy measures for continuous variables, either on a latitude-longitude grid or at an individual point, attempt to address questions like, "was the intensity of precipitation correctly forecast?" There are analogous questions that can be posed for core meteorological observables like the ones given below:

- Temperature [°C or °F]
- Dew-point Temperature [°C or °F]
- Humidity [%]
- pressureSurfaceLevel (local pressure at site elevation) [hPa, in. Hg]
- pressureSeaLevel (pressure extrapolated down to mean sea level) [hPa, in. Hg]
- solarGHI (solar energy input) [W m^{-2}, Btu $ft.^{-2}$]
- windSpeed [m s^{-1}, ft. s^{-1}]
- windGust [m s^{-1}, ft. s^{-1}]

2.3.1. Validation Statistics for Continuous Variables

For quantifying a model's performance with respect to continuous variables like temperature and pressure, a different set of statistical metrics are commonly used to characterize the strength of the relationship between forecasted (f_i) and observed (o_i) conditions (the number of forecast-observation pairs, N, is also important for the purpose of evaluating the statistical significance of any validation results). Thus, analysts typically use the following metrics to define continuous variable forecast accuracy for intensity/magnitude, rate, or accumulation attributes:

- Mean Error (ME) – The aggregate average difference between forecasts and observations, also known as Mean Bias:

$$ME = \frac{1}{N}\sum_{i=1}^{N} f_i - o_i$$

  A value close to 0 implies comparatively high forecast skill.

- Mean Absolute Error (MAE) – The average absolute difference between forecast and observed values. Unlike ME, MAE does not allow for cancellation between errors of opposing sign in the aggregate sum.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |f_i - o_i|$$

A value close to 0 implies comparatively high forecast skill.

- Root-mean Square Error (RMSE) – The square root of the mean-squared error between forecast and observed conditions. Like the MAE, RMSE does not allow cancellation of opposite-signed errors in the aggregate sum.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (f_i - o_i)^2}$$

Values close to 0 indicate better performance, on average.

- Pearson Correlation Coefficient (r) – A measure of the linear association between forecast and observed conditions; the correlation coefficient provides a general sense of how closely one variable will track with another.

$$r = \frac{\sum_{i=1}^{N} \left(f_i - \underline{f}\right)\left(o_i - \underline{o}\right)}{\sqrt{\sum_{i=1}^{N} \left(f_i - \underline{f}\right)^2}\sqrt{\sum_{i=1}^{N} \left(o_i - \underline{o}\right)^2}}$$

Here, the values of f and o with overbars indicate the sample mean value of forecasts and observations, respectively. An r value of 1 indicates perfect association between the sample values of f_i and o_i, whereas an r value of -1 suggests perfect "inverse" association where increases in f_i always imply decreases in o_i, or vice versa. Functions to compute r are readily provided in standard statistics software packages and they often include an estimate of the statistical significance of the resulting correlation. Statistical significance refers to the likelihood of the correlation being random chance. For example, one can imagine that the forecast and observation data could be re-sampled many, many times over and that r can be calculated during each pass. If r is close to either 1 or -1 (i.e., not 0) *and* the variation in the re-sampled collection of r values is small, then the result is taken to be statistically significant. As mentioned previously, the quantification of statistical significance will depend on the number of entries, N, in a given data sample.

2.3.2. Resistant Statistics for Non-normal Data

When forecast and/or observed samples contain large outliers (data points that are very *unlike* the rest of the sample) the resulting summary statistics (as defined above) can be wildly distorted (*Wilks*, 2016). The distortion in summary statistics results from outlier values being effectively "amplified" after raising them to a power greater than 1 and therefore overshadowing the contributions from possibly more numerous, yet smaller-magnitude data pairs in error computations. This may be the case when the underlying distribution of data does not resemble the Gaussian or "normal" distribution, as is commonly assumed. A set of summary statistics are available to mitigate the confounding influence of outlier data points for non-normally distributed data (i.e., non-Gaussian data or data not resembling the "Bell Curve").

- Median Absolute Deviation (MAD) – For when the sample distribution is skewed or non-normal. As with the mean statistics, it is valuable to look at the central tendencies of the skewed distribution.

$$MAD = median\{|f_i - o_i|\}$$

That is, MAD is the median of the set of absolute differences between forecast and observed values. Taking the median reduces the sensitivity of the resulting summary statistic to outlier data points. For a perfect forecast, MAD = 0.

- Spearman Rank Correlation (r_s) – After transforming the forecast and observed data from true magnitude to the respective rank in each sample, this calculation follows the same form as the pearson correlation coefficient. In other words, r_s is the r computed on the relative ranks of the underlying data inputs, 1 to N. The reader can find more information on how to calculate r_s in *Brown et al.* (2020) or use functions which are readily provided in scientific computing software packages.

On occasion, the analyst may be interested in understanding the "whole" behavior of skewed/non-normal data, i.e., characterizing the data's central behavior as well as points that occupy the extremes or "tails" of the sample distribution. Computing percentiles of forecast and observed data separately for comparison may be useful. One can display the probability density function (PDF; the distribution of values across the data range), then compare the cumulative distribution functions (CDF; the relative fraction/percentage of data that are less in magnitude than a specified threshold value), or finally examine a quantile-quantile plot of forecast vs. observed percentiles (see examples shown in *Wilks*, 2016).

2.3.3. Accommodating High-resolution Data and Volatility

An issue with using scalar metrics, like MAE (see **Sec. 2.3.1**), for tabulating errors is that they assume exact matching between model forecasts and observations, even though forecasts and observations may have different resolution/sensitivity. When a weather model resolves fine-scale and quickly evolving features such as wind gusts as well as localized temperature and moisture fluctuations, the exact timing of these phenomena in the forecast will potentially be out of sync with *in-situ* observations. Thus, using a metric like MAE for validation can create a "double-penalty" in cases of such a misalignment (*Wilks*, 2016); a model forecast of a "no" condition is penalized once where observations indicate "yes", and then the forecast is penalized once more for offsetting the "yes" condition to a location where observations indicate "no" accordingly (*WMO*, 2015).

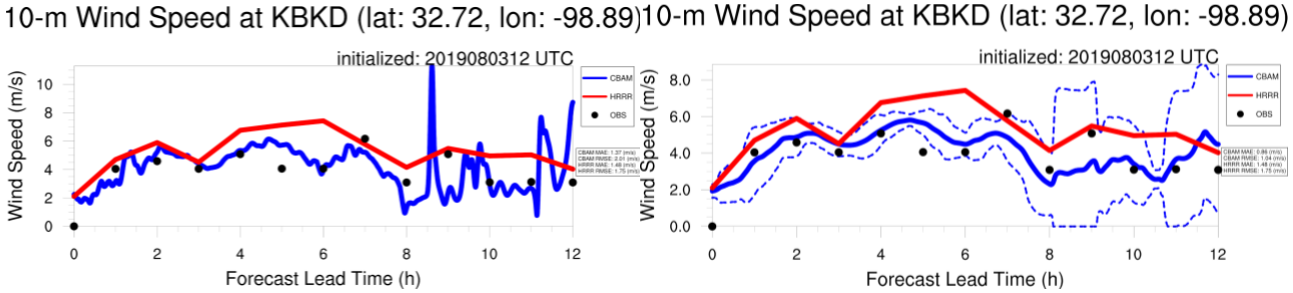

Figure 2. Wind speed forecasts at 10-m for a single observation location showing the contender forecast (CBAM) and comparison (HRRR) models compared to quality-controlled observations: incorporating high-frequency temporal variability from CBAM (left) and time-averaged CBAM forecasts with confidence intervals indicated by thin, dashed lines (right).

One basic approach to address misalignment between model forecast and observation resolution is to summarize forecast output over a window in time, effectively reducing model "volatility" to a mean plus standard deviation or a median and inter-quartile range. As suggested by example results in **Figure 2**, rapid fluctuations in forecast wind

speed differ slightly in time when compared with quality-controlled baseline observations. Temporal "windowing" renders central forecast values more in-line with observations and nearly all observed instances are captured within the model's standard deviation range, thus reducing the potential of involving "double-penalties" in validation analyses. Moreover, knowledge of the forecast's variability may also be of interest to the analyst and the aforementioned approach preserves such information.

3. Example Validation Studies

3.1 Case Study for Discrete Variables – NowCast Validation

Verification was performed with four months of Tomorrow.io's NC forecasts (July–October 2019), using NCAR MET during the assessment. The verification dataset was NOAA's/National Centers for Environmental Prediction's Stage-IV precipitation data, which has a 4-km grid spacing. The validation baseline data is independent from the Multi-Radar, Multi-Sensor precipitation product (the NOAA radar dataset used as input for Tomorrow.io's NC). Because Stage-IV data are hourly, the top-of-the-hour output was used to validate NC forecasts (e.g., the NC forecasts run at 12, 13, 14 UTC, etc.). Five thresholds were used based on observed one-hour accumulated precipitation: ≥ 0.01 in. (0.25 mm), ≥ 0.10 in., ≥ 0.25 in. (6.35 mm), ≥ 0.50 in. (12.7 mm) and ≥ 1.0 in. (25.4 mm).

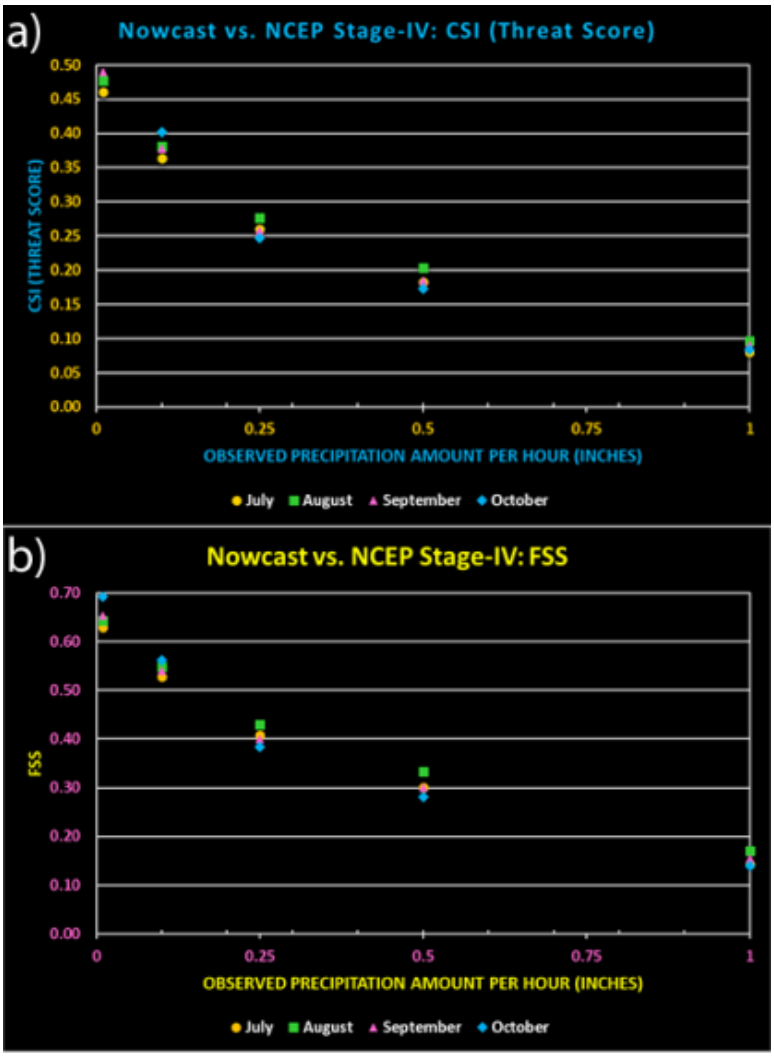

Figure 3. NowCast hour 1 forecast (a) CSI and (b) FSS over the CONUS for five one-hour precipitation thresholds (≥ 0.01 in. (0.25 mm), ≥ 0.10 in. (2.54 mm), ≥ 0.25 in. (6.35 mm), ≥ 0.50 in. (12.7 mm), and ≥ 1.0 in. (25.4 mm)), validated using NCEP Stage-IV data. Scores are shown for July (orange), August (green), September (pink), and October (blue) 2019.

The NC and Stage-IV datasets have slightly different nominal grid spacing, so pre-processing steps involved using NCAR MET to re-grid the Stage-IV data to match the NC grid spacing. Moreover, one minor issue arose from the fact that NC produces instantaneous precipitation rates, while Stage-IV produces hourly accumulations. To resolve this issue, the average hourly precipitation rate from NC (estimated once every six minutes) was used to calculate accumulated precipitation estimates. We used two common forecast verification skill scores: the commonly used CSI and the Fractions Skill Score (i.e., FSS; to be described below in **Sec. 4.2** below). In short, FSS is a spatial verification metric that does not penalize forecasts for close misses in space and time (e.g., the "double penalty").

Figure 3 shows CSI **(Fig. 3a)** and FSS (Fig. 3b) results for the continental United States domain during July–October 2019. As is typical, both skill scores decrease with increasing precipitation accumulation threshold. The results are also reasonably similar for all four months of the evaluation, suggestive of NC's consistent performance. For the ≥ 0.01 in. (0.25 mm) precipitation threshold, NC exhibits CSI values near 0.5 and FSS values greater than 0.6. Based on previous literature (*Roberts*, 2008; *Roberts and Lean*, 2008; *Mittermaier and Roberts*, 2010) and NOAA Weather Prediction Center verification statistics, these results suggest excellent performance for the question of "will precipitation occur?". While the skill scores for the ≥ 0.10 in. (2.54 mm) are slightly lower, CSI and FSS both indicate useful forecasts. Another interesting result is that the FSS (Fig. 3b) for the ≥ 0.01 in. (0.25 mm) precipitation threshold improved with each month, peaking around 0.7 during October. While higher scores are to be expected during the cool season with less sudden convective development and dissipation, the October results are still exceptional.

3.2 Case Study for Continuous Variables – CBAM Validation

A two-year, retrospective analysis was performed over the greater Toronto region in Canada to validate temperature, precipitation, and wind speed forecasts from CBAM against data from multiple ASOS sites between Canada and the United States. The station locations and their four-character identifiers are as follows:

- CWWZ = Port Weller Airport, Canada (latitude: 43.25, longitude: -79.21)
- CXET = Egbert Cs Airport, Canada (latitude: 44.23, longitude: -79.78)
- CXVN = Vineland, Ontario, Canada (latitude: 43.19, longitude: -79.39)
- CXTO = Toronto City Airport, Canada (latitude: 43.67, longitude: -79.40)
- KIAG = Niagara Falls International Airport, USA (latitude: 43.19, longitude: -79.21)

Note that the CBAM simulations were run at a resolution that was high enough to explicitly resolve fine-scale weather phenomena such as individual precipitation cells, rapid temperature fluctuations, and terrain-influenced flows/wind gusts. The PDF's of precipitation, temperature, and wind gusts are shown in **Figures 4-6** below.

Initial qualitative assessment shows that the CBAM model captures the climatology of precipitation reasonably well. There is a small tendency for CBAM to overestimate the frequency of low accumulation and underestimate the frequency of greater accumulation relative to observations. From **Figure 4**, it is also apparent that CBAM captures aspects of local precipitation variability well.

For temperature **(e.g., Fig. 5)** CBAM correctly captured bi-modal distribution characteristics at the four stations of interest, which is indicative of seasonal (e.g., summer vs. winter) variability. For the Port Weller Airport (station ID: CWWZ), CBAM handled the seasonal variations shown in observations, as evidenced by correct double-peak behavior. However, the forecast model tended to overestimate the frequency of moderate temperatures and underestimate the frequency of temperatures near 10-12°C as well as in excess of roughly 25°C; it remains to be seen whether the disagreement is due to local siting error or routine bias in the forecast model output. CBAM does well to represent the peak frequency of low to moderate wind gust events across all stations analyzed **(Fig. 6)**, but

small biases are noted as well as underestimation of a few extreme wind gusts at two sites, namely Niagara Falls International Airport (KIAG, United States) and Vineland, Ontario (CXVN, Canada).

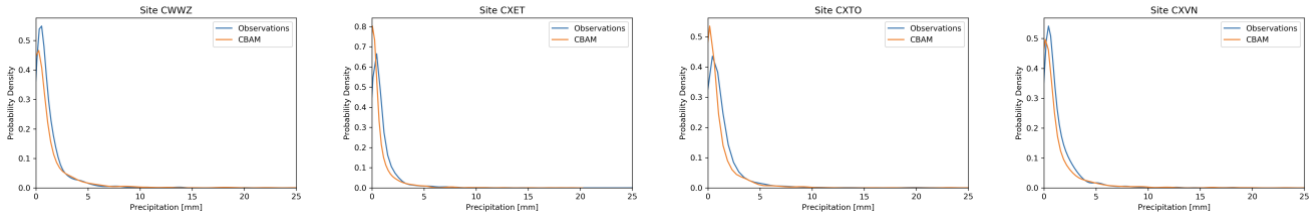

Figure 4. Probability density functions for two years of CBAM historical precipitation data compared to observations from four station observation sites in the greater Toronto region: CWWZ (left panel), CXET (middle left panel), CXTO (middle right panel), and CXVN (right panel). For each site, a total of approximately 18,000 observations were analyzed as part of the two-year validation.

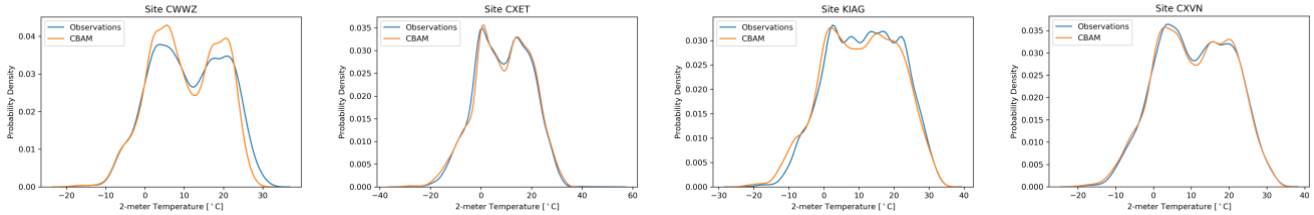

Figure 5. As in Fig. 4, but for 2-meter temperature data compared to observations for four surface observation sites: CWWZ (left panel), CXET (middle left panel), KIAG (middle right panel), and CXVN (right panel).

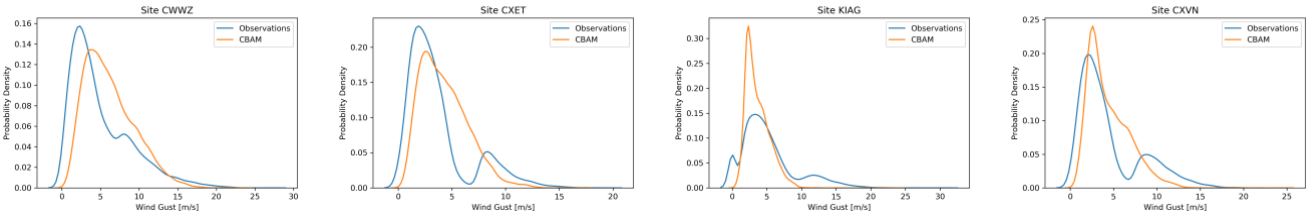

Figure 6. As in Fig. 4, but for 10-meter wind gust data compared to observations for four surface observation sites: CWWZ (left panel), CXET (middle left panel), KIAG (middle right panel), and CXVN (right panel).

Next, co-locating CBAM forecasts with ASOS data in time allowed for paired timeseries and summary statistics to be compiled for the two-year validation sample (see **Tables 1-3** below). CBAM forecast performance was evaluated according to the mean error ("Bias", ME), root-mean squared error (RMSE), mean absolute error (MAE), and bivariate (Pearson) correlation.

Table 1. Derived performance statistics from 2-years of CBAM historical one-hour precipitation forecasts compared to observations.

Site	Bias (ME) [mm]	Mean Absolute Error (MAE) [mm]	Root-mean squared Error (RMSE) [mm]	Bivariate Correlation (r)
CWWZ	-0.037	0.126	0.771	0.356
CXET	-0.0280	0.148	0.855	0.264
CXVN	-0.036	0.135	0.857	0.300
CXTO	-0.021	0.137	0.998	0.299

Table 2. As in Table 1, for CBAM's historical 2-m temperature forecasts.

Site	Bias [°C]	Mean Absolute Error (MAE) [°C]	Root-mean squared Error (RMSE) [°C]	Bivariate Correlation (r)
CWWZ	0.734	1.379	1.750	0.972
CXET	-0.199	1.378	1.950	0.971
KIAG	0.870	1.515	1.755	0.951
CXVN	0.1461	1.221	1.66	0.965

Table 3. As in Table 1, for CBAM's historical 10-m wind gust forecasts.

Site	Bias [m s^{-1}]	Mean Absolute Error (MAE) [m s^{-1}]	Root-mean squared Error (RMSE) [m s^{-1}]	Bivariate Correlation (r)
CWWZ	-1.128	2.214	2.664	0.709
CXET	-0.632	1.628	2.013	0.724
KIAG	1.438	2.096	2.725	0.775
CXVN	0.074	1.675	2.284	0.780

Examination of the PDF's shown in **Figs. 4-6** together with the scalar performance statistics shown in **Tables 1-3** above suggests that there are no consistent ME's/biases in the model forecasts between observation sites. Furthermore, the results suggest that MAE and RMSE statistics (i.e., average error magnitude) are relatively small compared to the mean values overall (<25-50% of the mean for each variable of interest). The correlation between forecasts and observations are strongest for 2-m temperature ($r > 0.95$), followed by 10-m wind gust ($r > 0.70$), and finally for one-hour precipitation accumulation ($r > 0.26$).

4. Advanced Validation Techniques

4.1 Cross Validation

As it pertains to validation practice, the input data are an important design consideration for every modeling system; if a given operational product ingests all available observational data, then there are no true sources of independent observations to serve as "ground truth" for validation. Thus, one approach is to develop a product such that observational inputs can be included or withheld at will. Cross validation is a technique where the same forecast is made repeatedly, holding out a different set of observations each time. This approach affords the analyst with a test dataset that can then be used in a fair comparison of the ensuing forecasts against ground truth observations (that which were withheld from a given forecast). This technique can also be applied retrospectively, whereby forecast data from different model versions can be compared, e.g., to track forecast improvement over time as new/different input data are ingested during production (see discussion in *Wilks*, 2016).

4.2. Neighborhood Methods for Validation

Validation methods to handle either discrete and continuous variables were discussed in Sec. 2 above. It was noted that inaccuracies in contingency or scalar error analyses can arise due to spatiotemporal mismatch (e.g., the "double penalty") when comparatively high-resolution model forecast data are involved. Thus, additional techniques have

been devised in order to potentially remedy mismatch issues during validation; one class of emergent validation techniques for this purpose uses "neighborhoods" surrounding the grid point (or grid points) of interest (e.g., *Gilleland et al.*, 2009). In general, the technique involves identifying a collection of adjacent grid points around the target point and subsequently tabulating the fraction of those neighboring grid points that satisfy a given criterion for the forecast and the observation separately. As an example, the test criterion could be the occurrence of precipitation accumulation larger than some threshold value. A related accuracy metric is the Fractions Skill Score (e.g., *Roberts and Lean*, 2008), which is defined as follows:

- Fractions Skill Score (FSS) – The relative difference between forecast probability, P_f, and observed probability, P_o, of event occurrence, aggregated over all M neighborhoods in the domain (*Wilks*, 2016).

$$FSS = 1 - \frac{\sum_{m=1}^{M} \left(P_{f,m} - P_{o,m}\right)^2}{\sum_{m=1}^{M} {P_{f,m}}^2 + \sum_{m=1}^{M} {P_{o,m}}^2}$$

  P_f and P_o are evaluated as the fraction of neighboring points that satisfy the test criterion for the forecast and observation data, respectively. If the forecast and observation fields closely resemble one another, then the difference between P_f and P_o will be small. Therefore, FSS close to 1 indicates high forecast accuracy.

In practice, the size of the neighborhood window can be varied during successive evaluations of FSS in validation analyses. Inspection of the variations in FSS with respect to neighborhood size can reveal insight about the scale of the greatest mismatch/displacement error in a paired forecast-observation data sample (*Roberts and Lean*, 2008).

5. Human-Observation-Only Situations

There are rigorous standards for atmospheric observation and every effort should be dedicated toward adhering to community-adopted best practices. In making observations, humans consider a limited spatial area (e.g., "outside the window" or "at a given location on the property") or a single moment in time, often without regard for the degree to which their own subjective observation represents true meteorological conditions in the immediate environment. Since it is very difficult to treat the various sources of uncertainty inherent in human-only observations, it is important to report weather conditions or parameters based on either directly or indirect measurement using instruments. A number of measurements serve as verifiable alternatives to human-only observations, such as using ceilometers (bottom-up approach) and satellites (top-down approach) to determine cloudiness/sky-cover fraction. Other examples are empirical formulas which combine temperature, moisture, and wind speed inputs to determine heat or wind-chill indices. With direct or indirect meteorological measurements collected and subsequently formatted for analysis, it is then possible to employ discrete and continuous variable validation methods (refer to **Secs. 2** above).

References

Bradley, A.A., S.S. Schwartz, and T. Hashino, 2008: Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score. *Weather and Forecasting*, 23, 992-1006.

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

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.

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

Mittermaier, M., and Roberts, N., 2010: Intercomparison of Spatial Forecast Verification Methods: Identifying Skillful Spatial Scales Using the Fractions Skill Score, *Wea. and Forecast.*, 25(1), 343-354. Retrieved Mar 4, 2021, from https://journals.ametsoc.org/view/journals/wefo/25/1/2009waf2222260

Roberts, N., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteor. Appl.*, 15, 163–169.

Roberts, N., and H. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, 136, 78–97.

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.

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