Generated using the official AMS LaTeX template v6.1

Interpretation of Probabilistic Surface Ozone Forecasts: A Case Study for Philadelphia

Nikolay Balashov,[a,b] Amy K. Huff,[c] and Anne M. Thompson[d]

[a] *NASA Global Modeling and Assimilation Office (GMAO), NASA Goddard Space Flight Center, Greenbelt, MD, 20771, USA*

[b] *Assistant Research Scientist, ESSIC, University of Maryland, 5825 University Research Court, College Park, MD, 20740, USA*

[c] *I.M. Systems Group, College Park, MD 20740*

[d] *Earth Sciences Division, NASA Goddard Space Flight Center, Greenbelt, MD, 20771, USA and University of Maryland-Baltimore County, JCET, Baltimore, MD 21228*

Corresponding author: Nikolay Balashov, nikolay.v.balashov@nasa.gov

ABSTRACT: The use of probabilistic forecasting has been growing in a variety of disciplines because of its potential to emphasize the degree of uncertainty inherent in a prediction. Interpretation of probabilistic forecasts, however, is oftentimes difficult deterring users who may benefit from such forecasts. To encourage broader use of probabilistic forecasts in the field of air quality, a process for interpreting forecasts from a statistical probabilistic air quality surface ozone model REGiS is demonstrated. Four procedures to convert probabilistic to deterministic forecasts are explored for Philadelphia metropolitan area. These procedures calibrate the predicted probability of daily maximum 8-hour average ozone exceeding a standard value by 1) estimating climatological relative frequency, 2) establishing a probability of an exceedance threshold as 50%, 3) maximizing the threat score, and 4) determining the unit bias ratio. REGiS is trained using 2000-2011 ozone season (May 1 to September 30) data, calibrated using 2012-2014 data, and evaluated using 2015-2018 data. Assessment of the calibration data with the Pierce Skill Score suggests an exceedance threshold based on climatological relative frequency for the conversion from probabilistic to deterministic forecasts. Calibrated REGiS generally compares well to predictions from the US national air quality model and operational "expert" forecasts over the evaluation time period. For other probabilistic models and situations, different procedures of converting probabilistic to deterministic forecasts may be more beneficial. The methods presented in this paper represent an approach for operational air quality forecasters seeking to use probabilistic model output to support forecasts designed to protect public health.

SIGNIFICANCE STATEMENT: As probabilistic forecasting is becoming more prevalent in the field of air quality, the purpose of this article is to draw attention to the importance of defining a framework to accurately interpret these forecasts. This work shows that (1) probabilistic forecasts are potentially more useful to forecasters when converted into deterministic forecasts and (2) that some conversion methods are more skillful than others. It is recommended that if National Weather Service begins to produce probabilistic air quality products it implements some of the strategies presented herein to help with the interpretation of such forecasts.

1. Introduction

Both professional and public utilization of probabilistic forecasts in atmosphere-related disciplines have been growing over the past two decades (Roulston and Smith 2002; Du et al. 2015). Public precipitation forecasts are almost always reported as probabilities for a given time period (Weather prediction center 2021). For example, the Storm Prediction Center (SPC) predicts regional severe weather by contouring areas with different colors designating probabilities of severe storm occurrence (Schneider et al. 2009). Advances in computation recently led the U.S. National Centers for Environmental Prediction (NCEP) and the Fleet Numerical Meteorology and Oceanography Center (FNMOC) to join forces and generate sophisticated ensemble products for windwave heights prediction (Alves et al. 2013). The rapid increase in probabilistic forecasting prompted a desire to maximize the benefit of such products (Bröcker and Smith 2007).

The actual utility of probabilistic forecasting has been difficult to ascertain, however, because a variety of reports suggest that the issue of interpreting probabilistic forecasts is not a solved problem. For instance, Demeritt et al. (2010) identified that staff at 21 out of 24 flooding prediction centers in Europe were not comfortable using probabilistic predictions even though such data were already available from various appropriate ensembles. Joslyn et al. (2009) set up a 5-day precipitation forecast experiment showing that a visual presentation of probabilistic forecast matters, and a person may misinterpret a probabilistic forecast as deterministic if the forecast type is not clearly defined. Using a flood control game at a conference, Ramos et al. (2013) showed that in principle, probabilistic forecasts are more beneficial than deterministic ones because their description of uncertainty allows for more carefully thought-out decisions. However, to be effective, such a

procedure requires an understanding of what probabilistic products mean and practice in their utilization.

In this paper, potential interpretation issues that may arise with probabilistic models applied to ozone (O_3) air quality forecasting are investigated. Tropospheric O_3 is the primary component of photochemical smog. Because O_3 is harmful to public health and the environment, the United States (US) Environmental Protection Agency (EPA) regulates O_3 as one of six "criteria" pollutants (Beckett 1991; Lee et al. 1996; Avnery et al. 2011; EPA 2015). As such, EPA has established a daily National Ambient Air Quality Standard (NAAQS) for O_3, last updated in 2015, of 70 ppbv averaged over 8 hours. To help protect public health, state, local, and tribal governments issue operational O_3 air quality forecasts for most major metropolitan areas in the US via EPA's AirNow program. These operational O_3 forecasts are similar to temperature or precipitation forecasts issued by the National Weather Service (NWS), except that instead of maximum temperature or expected rainfall amount, the air quality forecasts alert the public to expected concentrations of ground-level O_3. Air quality forecasts are issued using EPA's color-coded Air Quality Index (AQI), shown in Figure 1.

AQI Categories of Health Concern	AQI Index Values	Range of MDA8 O_3 (ppbv)	Meaning
Good	0-50	0-54	Air quality is considered satisfactory, and air pollution poses little or no risk.
Moderate	51-100	55-70	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101-150	71-85	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151-200	86-105	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201-300	106-200	Health alert: everyone may experience more serious health effects.
Hazardous	301-500	>200	Health warning of emergency conditions. The entire population is more likely to be affected.

FIG. 1. Correspondence between the U.S. Air Quality Index (AQI), daily maximum 8-h average (MDA8) O_3, and their meaning to the public. AQI can also be modified by other airborne pollutants; in this article only O_3 is considered.

Days when maximum daily 8-hour average O_3 (MDA8) is expected to exceed the NAAQS of 70 ppbv, termed "exceedance days", are of particular concern for operational air quality forecasters. An O_3 operational forecast of > 70 ppbv triggers an "Air Quality Alert" or "Action Day" in most jurisdictions, analogous to a Severe Thunderstorm Warning, which warns the public to take action to limit their outdoor exposure to potentially unhealthy levels of O_3. Deterministic air quality models

that predict concentrations of ground-level ozone in the future, such as the National Air Quality Forecasting Capability (Kang et al. 2005; Eder et al. 2006, 2009; Stajner et al. 2012; Chai et al. 2013; Lee et al. 2017; Campbell et al. 2022, NAQFC), are a vital tool to help operational forecasters identify exceedance days. Over the past 20 years, to provide even more skillful guidance for O_3 exceedance day forecasts, a probabilistic O_3 prediction framework has been widely recommended (Dabberdt et al. 2004; Djalalova et al. 2015; Hogrefe and Rao 2001; Kamińska 2018; Monache et al. 2008; Vautard et al. 2006). As computing power increases, air quality probabilistic models are expected to become more common, allowing for a wider application of machine learning and ensemble methods (Pinder et al. 2009; Garaud and Mallet 2011; Gilliam et al. 2015; Athira et al. 2018; Fanfarillo et al. 2019).

Despite the capacity provided by O_3 probabilistic forecasts, their wider adoption by the air quality forecasting community may be impeded by a challenge of how to interpret the output of probabilistic models in an operational setting (Demeritt et al. 2010; Joslyn et al. 2009). An example of probabilistic MDA8 O_3 forecast guidance is shown in Figure 2, using the Regression in Self Organizing Map (REGiS) statistical probabilistic model (Balashov et al. 2017). The model generates an occurrence probability that MDA8 will fall in different AQI categories: 0.17% Good (green), 69.19% Moderate (yellow), 30.55% Unhealthy for Sensitive Groups (USG; orange) and 0.09% Unhealthy (red). In this example, there is a 30.64% probability that an O_3 exceedance will occur (orange + red shading). For an operational forecaster, predicting if an O_3 exceedance will occur is of more value than predicting an exact MDA8 numerical value (Garner and Thompson 2013). The issues the forecaster faces are illustrated in Figure 2. Presented with the probabilistic guidance shown in Figure 2, does an operational forecaster predict an exceedance (Mason 1979; Wilks 2011)? What is the most skillful way to interpret probabilistic air quality model guidance in an operational environment?

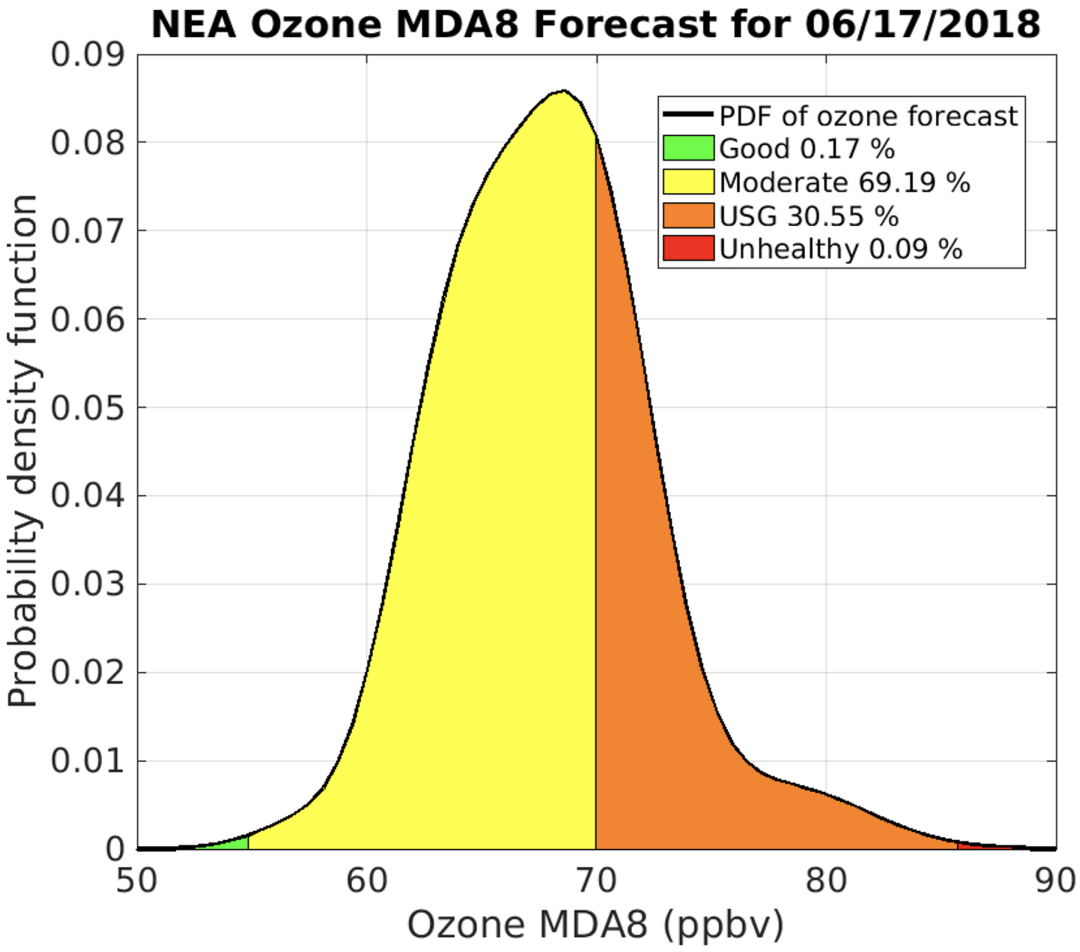


FIG. 2. An example of an experimental REGiS output showing a probabilistic O_3 forecast at the location of the Northeast Airport (NEA) regulatory O_3 monitoring site in Philadelphia, PA, valid on 06/17/2018.

The aim of this paper is to address these questions for operational O_3 forecasts in the Philadelphia metropolitan area of southeastern Pennsylvania by replicating the process of real-time forecasting of surface O_3 with a probabilistic statistical air quality model REGiS serving as a proxy for other potential probabilistic air quality forecasting systems. Philadelphia is selected as a representative example of an air quality jurisdiction that is in nonattainment of the 2015 O_3 NAAQS. Five local regulatory O_3 monitors are used to train, calibrate, and evaluate the model. Additionally, data from a representative weather station and a variety of products from operational weather forecasting models are used as predictor variables to drive REGiS. Four procedures are examined to convert REGiS probabilistic O_3 forecasts to deterministic yes/no forecasts of MDA8 O_3 exceedance occurrence. These procedures are based on 1) the climatological relative frequency, 2) the more likely event,

3) the threat score, and 4) the bias ratio. Calibrated versions of REGiS are compared to each other as well as to persistence, an older version of NAQFC, and operational (human) forecasts, with the goal of understanding how to maximize the benefit from the probabilistic O_3 forecasts provided by REGiS. In turn, these results should be more broadly applicable to probabilistic air quality models in general.

2. Data and methodology

Ozone and meteorological data

Forecasting region and its O_3 monitors

To reduce the forecasting experiment to a manageable size, only one forecasting region is chosen – Philadelphia, which is comprised of the following five counties in the state of PA: Bucks, Chester, Delaware, Montgomery, and Philadelphia (Department of Environmental Protection 2021, Figure 2). With regard to O_3 pollution, Philadelphia is well representative of the Mid-Atlantic area (generally includes states of NY, PA, DE, NJ, MD, DC, VA, and WV) and historically has experienced a notable amount of O_3 exceedances (Ryan et al. 2000; Kleinman et al. 2004). To establish whether a day is an O_3 exceedance or not, regulatory monitors of a designated location are used. The siting of pollutant monitors for measuring the ambient concentrations of criteria pollutants, of which O_3 is one, is a complex process outlined in the Code of Federal Regulations, Title 40 (Protection of Environment), Appendix D to Part 58 - Network Design Criteria for Ambient Air Quality Monitoring (`https://www.govinfo.gov/app/details/CFR-2019-title40-vol6/CFR-2019-title40-vol6-part58-appD`). In this study, five long-term AQ regulatory monitoring stations in Philadelphia are utilized (Table 1). The air quality data are acquired from the EPA's Air Quality System (AQS) database (`http://www.epa.gov/ttn/airs/airsaqs/`). The exceedances are determined from a monitor with the highest MDA8 reading in the region on a given day. For example, if five monitors have O_3 MDA8 values of 62 ppbv, 65 ppbv, 64 ppbv, 73 ppbv, and 72 ppbv then 73 ppbv will be considered a representative MDA8 number for the region. All the O_3 values used in this research are MDA8 and are calculated for O_3 seasons of 2000-2018, which in Philadelphia last from May 1st until September 30th (MJJAS). The corresponding exceedances are shown in Figure 3.

Philadelphia (southeastern PA)				
Site Full Name	Site Acronym	Site Air Quality System (AQS) ID	Latitude (°N)	Longitude (°W)
Bristol	BRIS	420170012	40.12	-74.88
Chester	CHES	420450002	39.84	-75.37
North East Airport	NEA	421010024	40.08	-75.01
New Garden	NEWG	420290100	39.83	-75.77
Norristown	NORR	420910013	40.11	-75.31

TABLE 1. Surface O_3 regulatory monitoring stations for Philadelphia, PA used in this study.

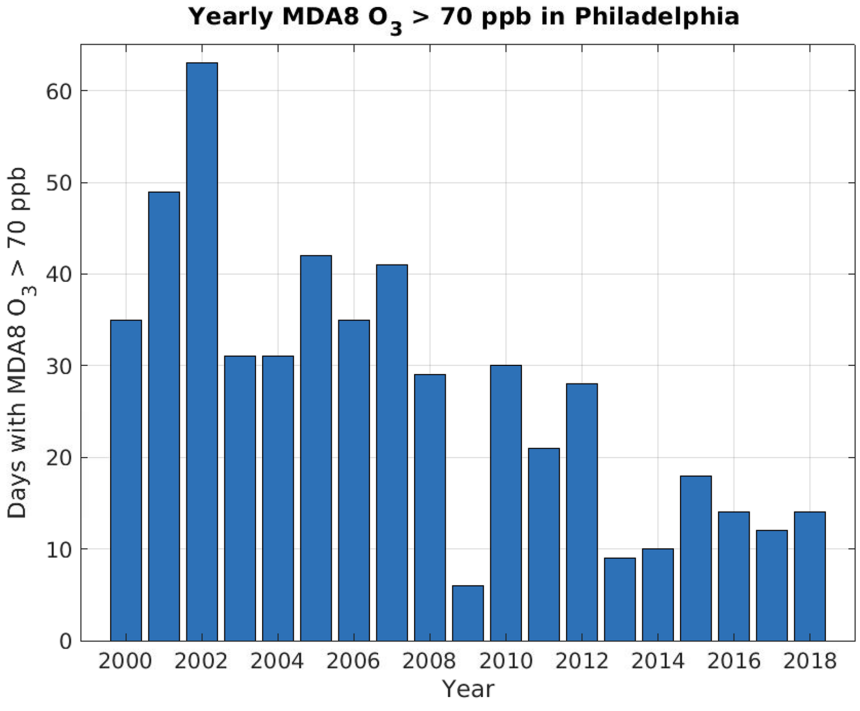


FIG. 3. Yearly number of days with MDA8 O_3 greater than 70 ppbv taken from the O_3 season (May 1st to September 30th) in the Philadelphia region over 2000-2018.

1) METEOROLOGICAL DATA

Meteorological data from Northeast Airport in Philadelphia County (KPNE; Figure 4c) are used as representative of conditions in the Philadelphia forecast area. The KPNE station is not representative of the extremely urbanized area of downtown Philadelphia, but rather was chosen to represent the weather in the entire O_3 forecast area encompassing Bucks, Chester, Delaware, Montgomery, and Philadelphia counties. The data were obtained from the National Climatic Data Center (NCDC). Synoptic spatial data used by REGiS come from the 2nd-generation NOAA Global Ensemble Forecast System (GEFS) Reforecast control run. GEFS Reforecast v2 is a dataset with historical weather forecasts; it is available at `http://www.esrl.noaa.gov/psd/forecasts/reforecast2/`. Prediction data for the meteorological regression predictors used in REGiS come from Global Forecast System (GFS) and the National American Mesoscale Forecast System (NAM) Model Output Statistics (Carter et al. 1989, MOS). The MOS method applies a regression model trained on the archived model data from GFS and NAM to adjust the station-specific meteorological forecasts from operational GFS and NAM. Current MOS data are available at `https://vlab.noaa.gov/web/mdl/mos-archived-data`. Archived MOS data from GFS and NAM models are used in REGiS to simulate realistic prediction of meteorological variables. The weather forecast is generated using 12 UTC product issued prior to the day of the forecast by averaging GFS and NAM forecasts for temperature (18 UTC and 21 UTC), relative humidity (18 UTC and 21 UTC), wind direction (15 UTC, 18 and 21 UTC), and wind speed (15 UTC, 18 and 21 UTC).

a. Basic overview of REGiS

The REGiS model is a statistical probabilistic model capable of predicting MDA8 O_3 and is described in detail in Balashov et al. (2017) with only essential features presented herein. REGiS uses Kohonen's Self-Organizing Maps (SOM), a spatial clustering procedure that is driven by unsupervised learning (Kohonen 2013), to identify a variety of weather regimes common to a geographic location using spatial meteorological variables. In the current study, two such spatial meteorological variables are used: 500 hPa geopotential height and 850 hPa temperature. Figure 4 shows two domains of REGIS: the outer domain (Figure 4a) is used for 500 hPa geopotential height and the inner domain (Figure 4b) is used for 850 hPa temperature. The number of regimes used in

the model's training phase is subjective and depends on how much training data are available. The map size used in the current version of REGiS is 7 by 8, identifying 56 weather regimes over the training period of 12 years starting from MJJAS 2000 and going until the end of MJJAS 2011.

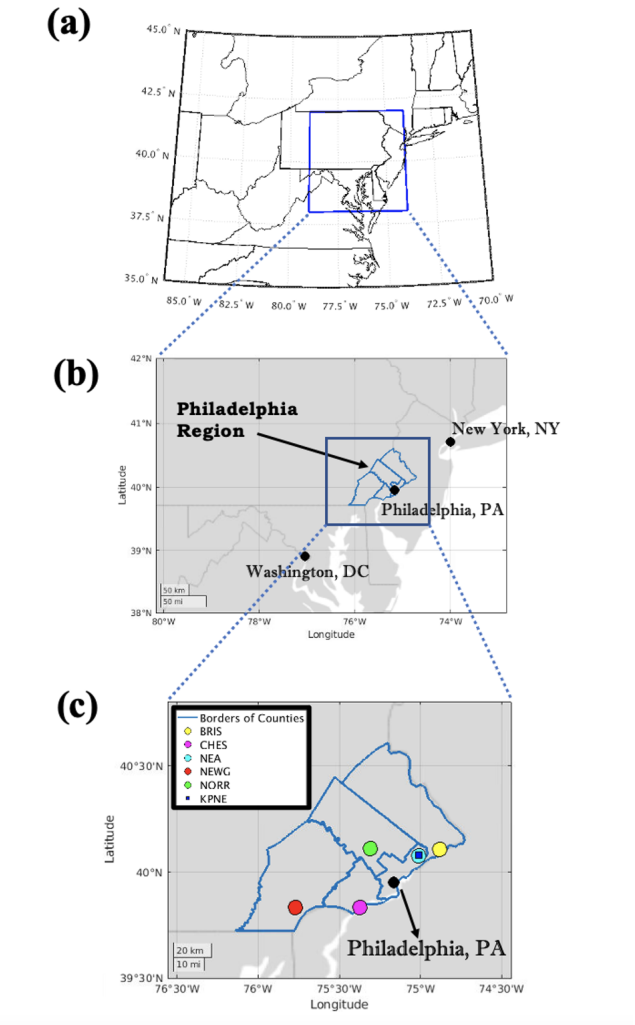


FIG. 4. **(a-b)** REGiS SOM domains for the Philadelphia airshed region, where **(a)** indicates inner domain used for the weather pattern classification by 850 hPa temperature and **(b)** shows outer domain that is used for the weather pattern classification by 500 hPa geopotential height. **(c)** Philadelphia region of interest showing the five air quality monitors and the weather station KPNE located at the airport.

Once weather regimes are determined, a separate stepwise weighted quadratic regression per surface O_3 monitoring station is created for each one of the regimes based on local meteorological variables such as temperature, relative humidity, wind speed, and wind direction measured by a

nearby Automated Surface Observing Systems (ASOS; in this study, KPNE) as well as on additional predictors of zenith angle and previous-day MDA8 O_3. Only the days that comprise a weather regime are used for the generation of a regression. The predictand of the mentioned regressions is MDA8 O_3.

The abovementioned procedures describe a training phase of the REGiS. The prediction phase is executed with the help of a publicly available NWP system. Such a system makes a prediction of spatial meteorological variables with a specific lead-time. Then REGiS identifies which one of the 56 weather regimes (given the current Philadelphia set-up) is the most like the NWP predicted spatial variables based on Euclidian distance metric. The regime that is the most like the predicted one receives the highest weight, whereas nearby regimes receive less weight and so on (the weight distribution depends on an algorithm chosen by a user; for more details see supplementary material and Balashov et al. (2017)). When all the relevant regimes are identified, the corresponding weighted regressions are solved using predicted meteorological variables from the NWP products and other relevant predictors to make a probabilistic MDA8 O_3 forecast (Figure 5). O_3 MDA8 exceedance probabilities are calculated at each of the sites separately, but because EPA guidelines count a reading from a site with the highest MDA8 value as representative, a monitor showing the highest probability of an exceedance in a region (in this case southeastern PA) is used to produce a daily forecast (Figure S5).

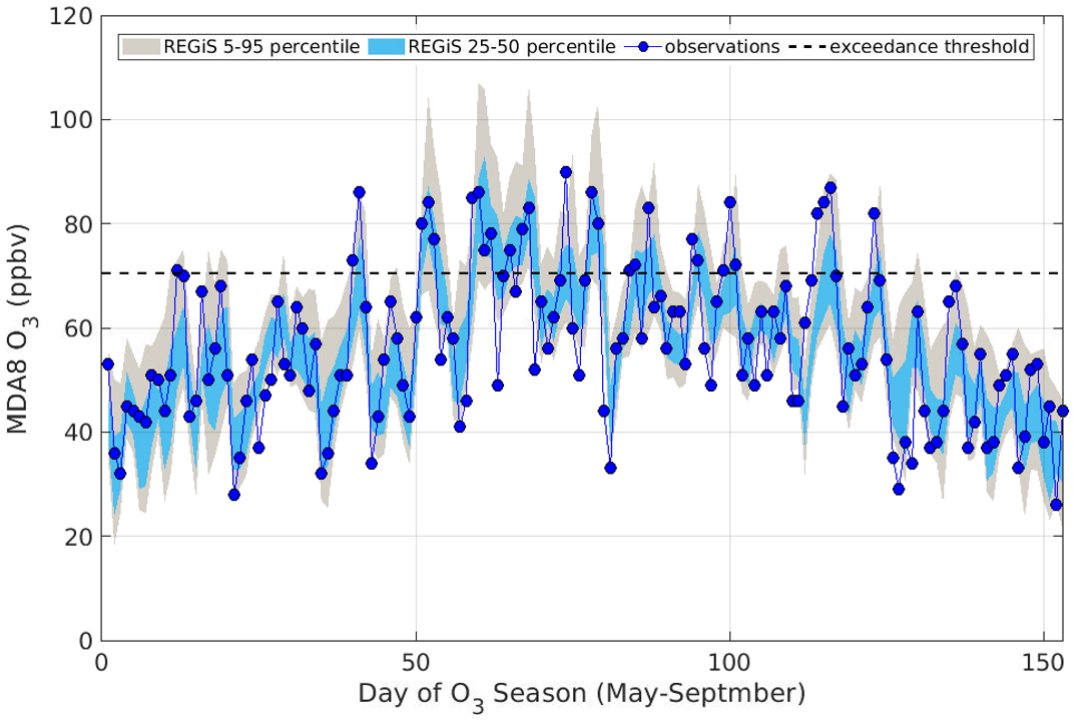


FIG. 5. Example time series of MDA8 O_3 REGiS probabilistic forecasts for 2012 MJJAS for Philadelphia. Light blue shading indicates 25-50 percentile region of the probabilistic forecast and grey shading indicates 5-95 percentile of the probabilistic forecast. Dark blue dotted line denotes MDA8 O_3 observations. Dashed black line shows EPA exceedance MDA8 O_3 standard of 70 ppbv.

b. The experiment

1) EXPERIMENTAL SETUP

A forecasting scenario is set up for REGiS to explore a variety of approaches that convert probabilistic forecast guidance to deterministic guidance. Operational forecasting environment is simulated for a hypothetical human forecaster working in the period 2015-2018. Therefore, observations prior to 2015 are available to train and calibrate REGiS, while data from 2015 to 2018 are used to evaluate the model.

Data from 2000 to 2011 are used to train REGiS to ensure enough data to drive the model, and to make sure that REGiS is not overly sensitive to higher MDA8 O_3 values, which have become relatively uncommon due to the 1997 NAAQS change (Lin et al. 2017). In the next step, the independent 2012-2018 dataset is partitioned into two parts: 2012-2014 is used for calibration and

2015-2018 is used for evaluation. The calibration procedure involves four different approaches: the climatological relative frequency, the more likely event, the threat score, and the bias ratio (Wilks 2011). These approaches are investigated in the search of the optimal probability threshold for converting probabilistic forecasts to deterministic forecasts. Then in the evaluation phase the optimal exceedance threshold of REGiS, determined in the calibration procedure, is compared against other forecasts, including persistence, NAQFC deterministic model output, and operational forecasts.

2) Exceedance probability thresholds

Four approaches examined here that establish exceedance probability thresholds are 1) the climatological relative frequency, 2) the more likely event, 3) the threat score, and 4) the bias ratio (Wilks 2011). These approaches define the way in which a REGiS probabilistic forecast will be considered as predicting an MDA8 O_3 exceedance or non-exceedance event. The climatological relative frequency method requires calculation of a climatology of exceedances and non-exceedances over the calibration time period. Then it is assumed that the calibration time period is representative of an evaluation time period. If a given probabilistic forecast predicts an exceedance with a higher chance than the calibrated climatology, then an exceedance is predicted. In the case of the more likely event methodology, calibration is not needed, and if a probabilistic forecast indicates an exceedance with over 50% certainty, then exceedance is forecasted; otherwise, it is a non-exceedance. To calculate threat score, a 2x2 contingency table is required (Figure 6). A contingency table displays counts of four different "yes"/"no" forecast/observation possibilities once the probabilistic forecast is converted into a deterministic yes/no forecast; here "yes" stands for an exceedance and "no" for a non-exceedance. In Figure 6, a represents a forecast "hit" (yes/yes), b is a "false alarm" (yes/no), c is a "missed forecast" (no/yes), and d is a "null forecast" (no/no) for forecast and observation, respectively. Once the contingency table is created, the threat score can be calculated as follows:

$$threat\ score = \frac{a}{a+b+c}. \tag{1}$$

The threat score is the ratio of correctly forecasted exceedances to all the other instances of forecasted and observed exceedances. It is desirable to maximize this ratio to ensure only the correct exceedances are forecasted. This means that during the calibration process, iterations occur

through all the possible probabilistic thresholds from 1-100, and the threshold that maximizes threat score is selected.

Lastly, the bias ratio is defined as follows:

$$bias\ ratio = \frac{a+b}{a+c}. \tag{2}$$

The definition can be interpreted as the ratio of forecasted exceedances to observed exceedances. The most skillful possible value for the bias ratio is 1. To determine at what threshold the bias ratio approaches 1, the same calibration procedure iteration as described for the threat score is applied.

		Observed: Yes	Observed: No
Forecasts	Yes	a (hit)	b (false alarm)
	No	c (miss)	d (correct rejection)

FIG. 6. A 2x2 contingency table indicating relationship between combinations of yes/no possibilities of forecast/observation counts shown by letters a through d. In this article "yes" means an O_3 exceedance occurred, while "no" means that there was no O_3 exceedance detected.

All the described thresholds can be visualized as a function of the hit rate and the false alarm rate using a Receiver Operating Characteristics (ROC) curve (Wilks 2011). The hit rate is defined as:

$$hit\ rate = \frac{a}{a+c}. \tag{3}$$

And the false alarm rate is expressed as:

$$false\ alarm\ rate = \frac{b}{b+d}. \tag{4}$$

The ROC curve is a common technique used to compare a variety of thresholds against each other with an assistance of the hit and false alarm rates metrics. The ideal threshold would produce

the hit rate of 1 and the false alarm rate of 0, but usually that is unattainable, so a compromise must be found. The ROC curve is useful for comparing a variety of thresholds based on hit and false alarm rates and is used in this study for this purpose; however, usually it provides subjective guidance regarding which threshold is more useful unless a number of hit and false alarm rates is of primary interest in a forecast. Here the Pierce Skill Score (PSS) is selected to assess the four thresholds described above with 2012-2014 calibration data, and subsequently the most skillful threshold based on the PSS is chosen for the 2015-2018 real-time forecasting evaluation.

The PSS is computed as:

$$PSS = \frac{ad - bc}{(a+c)(b+d)}. \tag{5}$$

The PSS has been shown to give an equitable score, allowing for a fair identification of which forecasting system is random and which is skillful in producing binary predictions (Hogan et al. 2010). While the inequitable skill scores tend to favor certain outcome events at the expense of other outcome events (Finley 1884; Gandin and Murphy 1992, e.g., few "yes" and many "no" outcomes). The PSS does not provide a complete evaluation of a forecast as it is designed to only deal with binary predictions that are generated by setting a subjective threshold. Therefore, an intensity of a pollution event is not evaluated when the PSS is used. Additionally, if there is a specific requirement from a binary prediction (e.g., minimizing false alarms) then it is possible that a metric other than the PSS could be more suitable.

c. NAQFC and operational MDA8 forecasts

NAQFC is a real-time air quality prediction system covering Continental United States (CONUS) domain capable of providing O_3 and $PM_{2.5}$ forecast guidance developed as a collaboration between EPA and NOAA. NAQFC has undergone a variety of changes since 2004 transitioning through several versions of Community Multi-Scale Air Quality (CMAQ) model driven by various meteorological models from NCEP's Eta to Finite-Volume Cubed-Sphere (FV3) in the GFS model that is currently used (Campbell et al. 2022). The O_3 forecasts from NAQFC utilized in this study come from an older version of NAQFC with the offline-coupled North American Mesoscale Model Forecast System on the B-Grid (NMMB) and CMAQv5.0-5.2 (Lee et al., 2017). Model outputs of the NAQFC 48-hour MDA8 O_3 forecasts at a resolution of 12 km were provided by the Philadel-

phia air quality operational forecasters for years 2015-2018 (National centers for environmental information 2022).

Operational MDA8 O_3 forecasts for Philadelphia represent an aggregate of expert knowledge with various air quality forecasting tools including NAQFC. These are the forecasts that were issued to the public daily during the study period. The forecasts used in this paper for the evaluation are acquired from air quality forecasters based at Pennsylvania State University, who issued the daily forecasts for the public.

3. Results and discussion

a. Determining exceedance probability thresholds

Using calibration data from 2012-2014, the ROC curve is graphed in Figure 7 to visualize all possible exceedance probability thresholds for REGiS as a function of the hit rate and the false alarm rate at MDA8 standard of 70 ppbv, the current daily O_3 NAAQS. On the ROC curve, four colored diamonds indicate the four exceedance thresholds based on the climatology, the more likely event, the threat score, and the bias ratio. Figure 7 demonstrates that choosing a threshold based on the ROC curve is subjective since a compromise must always be made between hit rate and false alarm rate, and it is not clear which decision objectively makes more sense. For example, in Figure 7 the climatology-based probability threshold has the best hit rate of the methods, but it also has the highest false alarm rate, so it may not necessarily be the optimal choice for setting the exceedance threshold. The ROC curve is useful in general assessment of probability thresholds, but the current case requires further specification to determine probability threshold for the 2015-2018 evaluation period and is thus supplemented with the PSS.

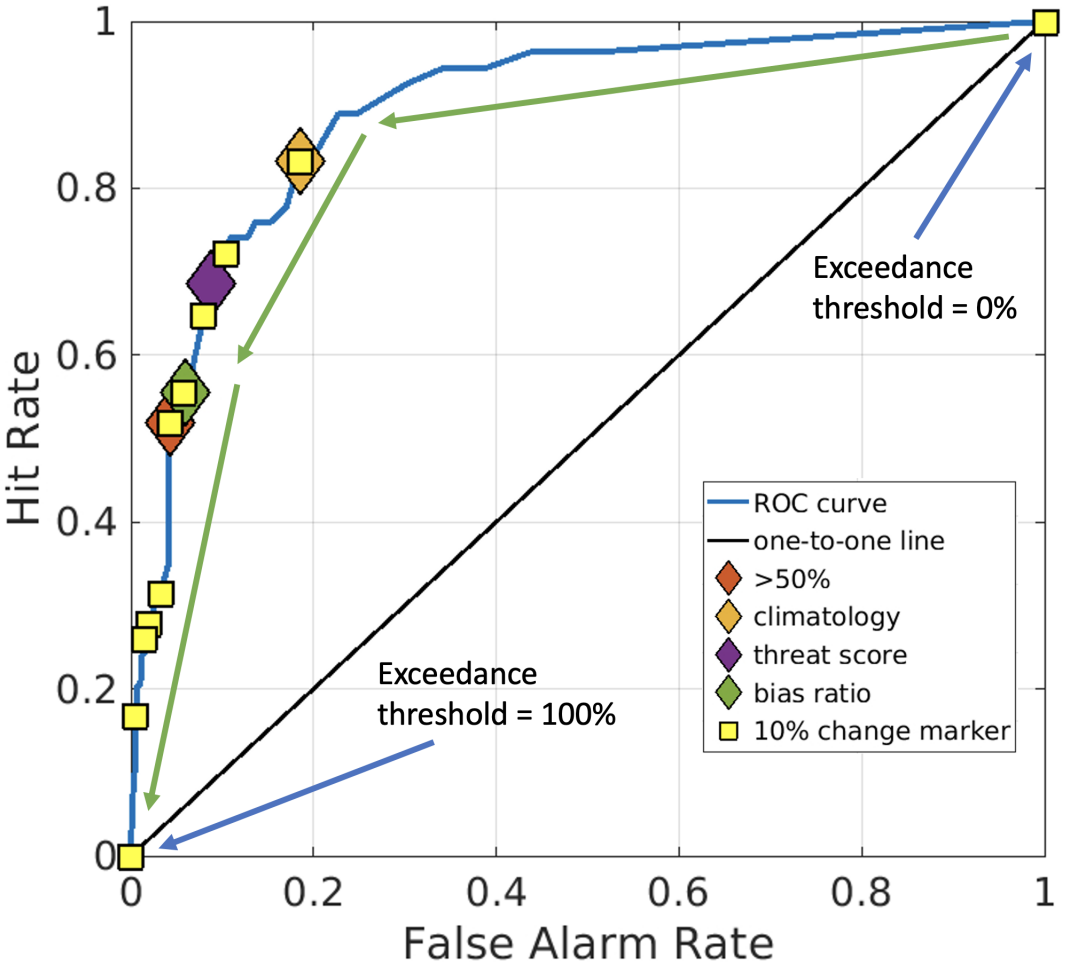


FIG. 7. ROC curve indicating REGiS exceedance probability threshold as a function of hit rate and false alarm rate assuming MDA8 O_3 standard of 70 ppbv using 2012-2014 calibration data. Yellow squares indicate exceedance probability thresholds every 10% starting from 0% in the top right and going to 100% in the bottom left as shown by the green arrows. The best forecast would have H of 1 and F of 0 (or as close as possible to that ideal). The colored diamonds are four different REGiS exceedance thresholds that are being assessed with the calibration data. Only one of those can be used for the real-time forecasting evaluation period of 2015-2018.

Even though the daily NAAQS for MDA8 O_3 is currently 70 ppbv, PSS is calculated for a variety of MDA8 O_3 standard values ranging from 75 ppbv to 45 ppbv. By using a sliding MDA8 O_3 standard, the performance of each approach can be approximated in different environments, and the one that produces the highest overall PSS can be selected. This calibration process is illustrated in Figure 8, where the PSS is evaluated for every standard and then averaged for every shown group.

For instance, taking the first group with 71-75 ppbv range, the PSSs are estimated separately for standards 71 ppbv, 72 ppbv, 73 ppbv, 74 ppbv, and 75 ppbv, and then averaged to get a representative group score.

Figure 8 demonstrates that the climatology-based approach is the optimal choice for the evaluation stage based on the PSS, as it performs more skillfully than the rest of the approaches in the first three MDA8 O_3 standard categories. The next best methodology is the threat score, on average performing almost as good as climatology. The more likely event and the bias ratio based approaches perform more skillfully when the O_3 standard decreases to 50-60 ppbv range, increasing the number of exceedances. However, when there are too many exceedances in the lowest category of 46-50 ppbv, the more likely event approach becomes noticeably worse again. As expected, persistence is the worst approach in all the categories. These examples show that varying exceedance values substantially affect the methods' performance. And because the number of exceedances that will occur in an operational forecasting setting is not known a priori, an educated guess must be made based on the available information. In this study, the calibration data of 2012-2014 are assumed to be representative of 2015-2018. Therefore, the climatology methodology is chosen for the operational environment evaluation.

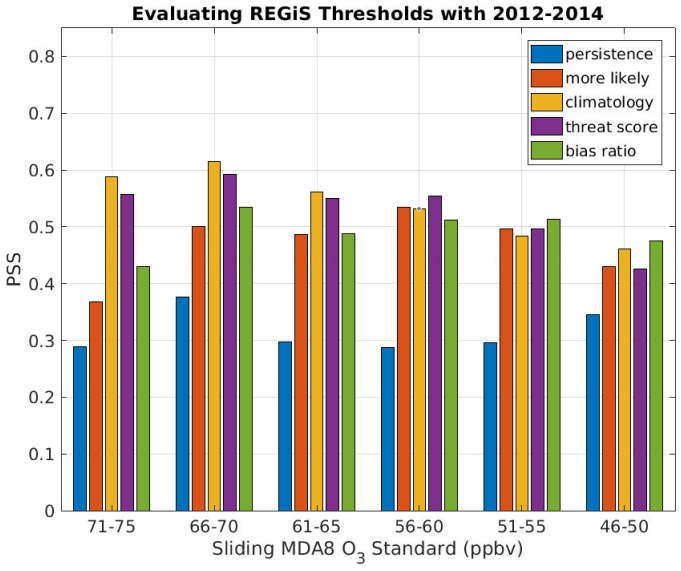


FIG. 8. Calibration of REGiS performed by evaluating four different exceedance probability thresholds and persistence using the PSS with 2012-2014 data for a variety of MDA8 O_3 standards ranging from 75 ppbv to 45 ppbv. The PSS is averaged over the shown standard intervals.

b. Operational environment simulation and corresponding evaluation

To verify that the climatology approach is indeed the most optimal choice in the current case, Figure 9 compares the PSSs of REGiS probabilistic forecasts using four different approaches for every year in the 2015-2018 period. Similar to what was observed in the calibration phase in Figure 8, climatology (in yellow) on average is the most skillful threshold across the range of MDA8 O_3 standards. The second-best threshold is based on the threat score (in purple). The PSS becomes more similar among the methodologies for all the years at about 50-55 ppbv MDA8 O_3 standard, when the number of the exceedances increases.

For the case illustrated (Philadelphia), the overall number of exceedances in the calibration phase (2012-2014) is comparable to what is observed in the evaluation phase (2015-2018); however, there is a noticeable noise from year to year (Figure 3). Under the assumption that the number of exceedances remains the same as observed during the calibration phase (as well as the connection between O_3 and meteorology used by REGiS to make predictions), the climatology-based threshold PSS should be somewhere around 0.6 during the evaluation phase of the 71-75 ppbv category. This value is approximately found in 2015 (Figure 9a) and 2018 (Figure 9d) with the values being 0.6 and 0.67. But for the years 2016 and 2017 the values are lower (0.4 and 0.46). Taking the years 2015-2018 together gives an average PSS of 0.53, which is below the calibration phase value of 0.6; nevertheless, the skill score of the climatology-based threshold is still the highest relative to the other three approaches for the 71-75 ppbv range. So, although the prediction conditions, such as variability in emissions or local meteorology may have changed since 2012-2014, the overall relationship remains correct for 2015-2018. This may not always be the case, and it is important to keep that in mind in case additional information regarding the operational forecasting period is available (e.g., introduction of new emission policies).

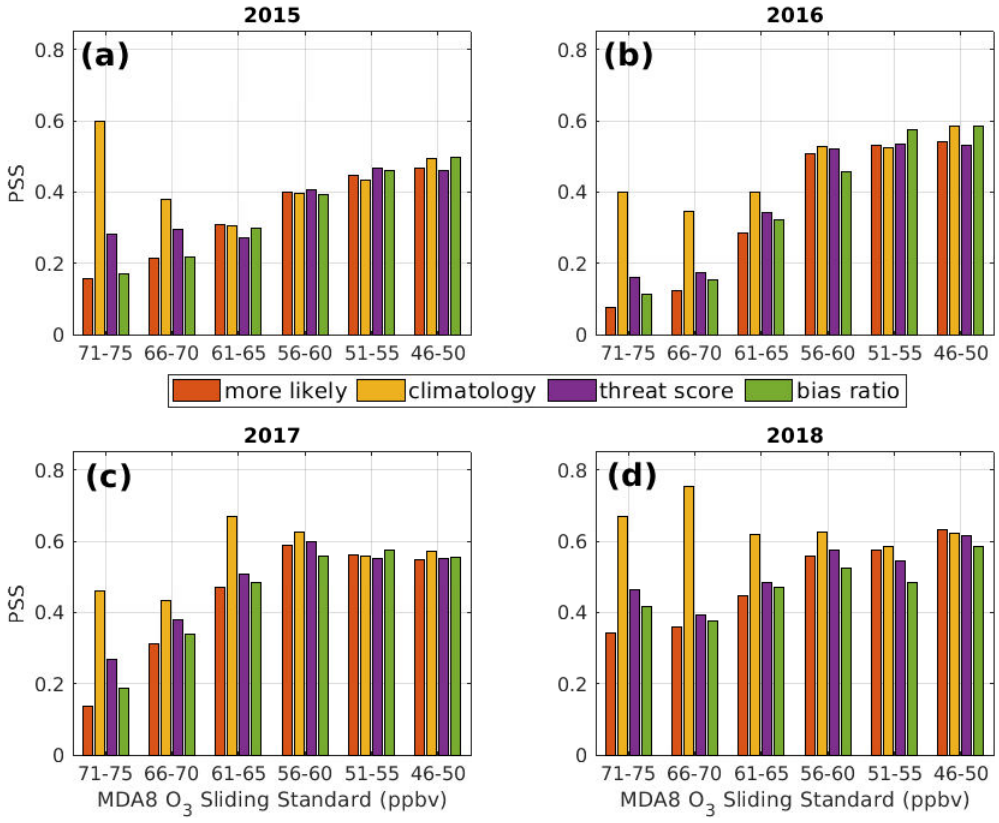


FIG. 9. Comparison of REGiS probabilistic forecasting skill (based on the PSS) with four different types of thresholds (discussed in section 3.1) for a variety of MDA8 O_3 standards using 2015-2018 independent data. The panels correspond to a different year: **(a)** 2015, **(b)** 2016, **(c)** 2017, and **(d)** 2018. The PSS is averaged over the shown standard intervals.

Now that it has been established that for the evaluation period of the current experiment climatology-based approach is indeed the preferred choice, according to the PSS statistic, comparison of REGiS against other forecasts can be performed. For the operational forecasting evaluation period, the climatology based REGiS is compared against persistence, NAQFC, and operational forecasters in Figure 10. Persistence, NAQFC model output, and operational forecasters only generate deterministic forecasts rather than probabilistic forecasts. Therefore, when they are evaluated against probabilistic REGiS forecasts, they must be classified into "yes" or "no" categories. For example, if the NAQFC predicts an MDA8 O_3 of 68 ppbv then it falls into the "no" category because it is below the 70 ppbv MDA8 O_3 standard.

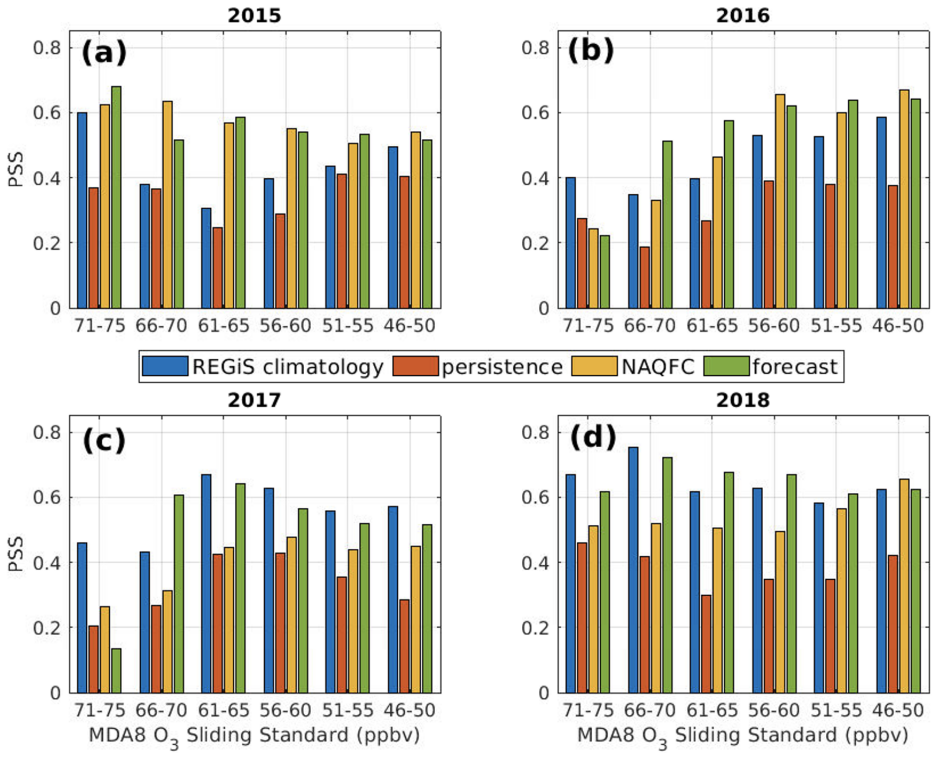


FIG. 10. Evaluation (using the PSS) of REGiS probabilistic forecasts with climatology-based thresholds against persistence, NOAA NAQFC model, and AQ forecasters for a variety of MDA8 O_3 standards using 2015-2018 independent data. The panels correspond to a different year: **(a)** 2015, **(b)** 2016, **(c)** 2017, and **(d)** 2018. The PSS is averaged over the shown standard intervals.

Overall, the highest PSS in Figure 10 belongs to the operational forecasters with REGiS and NAQFC being comparable to each other, while persistence shows the lowest prediction skill. On average, NAQFC performs more skillfully in 2015 (Figure 10a) and 2016 (Figure 10b), whereas REGiS is more skillful in 2017 (Figure 10c) and 2018 (Figure 10d). NAQFC that is evaluated in this work is driven by 2011 National Emission Inventory (NEI) and corresponding time-lag plays an important role in the uncertainty of deterministic NAQFC MDA8 O_3 predictions (NOAA Air Resources Laboratory 2019). REGiS is not necessarily a more skillful forecasting system than NAQFC and neither it is the goal of this article to evaluate REGiS, but due to a variety of complexities that require successful performance of NAQFC (e.g. emissions, transport, chemistry), REGiS can sometimes outperform it.

The above analysis demonstrates the advantages of calibrating probabilistic forecasting models for dichotomous events. Without the calibration procedure, it is not clear how to apply REGiS in an operational setting and how to evaluate the model against other forecasting tools. The precise process of reducing a probabilistic forecast to "yes" and "no" allows for an opportunity to make a clear decision and quantification regarding prediction of an MDA8 O_3 exceedance. Figure 10 indicates that in an independent operational forecasting scenario, calibrated REGiS performs on par with NAQFC and considerably more skillfully than persistence, demonstrating that REGiS can be a useful tool in an arsenal of an air quality forecaster.

4. Conclusion

Although probabilistic O_3 MDA8 forecasts hold more information than deterministic forecasts they pose an interpretation challenge to air quality forecasters. The NAAQSs are threshold values and do not lend themselves to probabilistic communication to the public. The O_3 forecaster benefits from the deterministic forecast guidance because they must issue a deterministic forecast of yes/no on Air Quality Alert to the public. This is similar to tornado forecasting, where there is a value in probabilistic information prior to the event, but when issuing a Tornado Warning, it is a yes/no proposition.

This article suggests that it is beneficial to perform a calibration procedure, converting a probabilistic forecast to yes/no prediction, on any type of probabilistic air quality model to maximize the usefulness of the product to an air quality forecaster. Such a process is thoroughly documented in the current work using a statistical probabilistic surface O_3 model called REGiS, which uses a combination of regression and SOMs to generate a one-day lead prediction of MDA8 O_3 in the form of a PDF. To implement a calibration, the probabilistic forecast is converted to a binary forecast by determining an exceedance probability threshold using four approaches: more likely event, the climatological relative frequency, the threat score, and the bias ratio. These approaches are assessed with the PSS. Results indicate that the forecasts provided by calibrated REGiS in the form of “yes” or “no” can be generally as good as the forecasts provided by NAQFC or better, underlining the value that could be extracted from the probabilistic forecasts of REGiS. For this study the climatology-based approach outperforms others when evaluated with the PSS, but this may not always be the case for a different model system or prediction environment.

Another potential way of selecting the exceedance thresholds would be to perform a value and cost-loss ratio analysis as a function of exceedance probability threshold as exemplified in Garner and Thompson (2013) article for the Baltimore, MD, region. However, such a procedure is generally more complex for a model development group to accomplish and would likely require a guidance from a policy maker. Hence, calibrating probabilistic forecasts as described in this article is presently the most direct way to increase the value of probabilistic MDA8 O_3 forecasts provided to professional forecasters.

Acknowledgments. This project started at Pennsylvania State University (PSU) as a post-PhD investigation, supported by NASA's Applied Sciences Air Quality Science Team (funding to Anne Thompson) and received considerable input from George S. Young, Steven Greybush, Benjamin Shaby, and William Ryan. Many of the project's simulations and calculations were performed using clusters maintained by the PSU Department of Meteorology and Atmospheric Science. This work was partially financially supported by Kenneth Davis at PSU, a NASA Postdoctoral Program (NPP) appointment with Universities Space Research Association (USRA), and an appointment with Earth System Science Interdisciplinary Center (ESSIC) of University of Maryland (UMD), while Anne Thompson was part of the NASA Health and Air Quality Applied Sciences Team (HAQAST). Special thanks to Lesley Ott at NASA's Global Modeling and Assimilation Office (GMAO) for support of this study. We also thank the Pennsylvania Department of Environmental Protection (DEP) for partial support of Amy Huff's contributions to this project. Much gratitude to the three anonymous reviewers who helped us to substantially improve the quality of this manuscript.

Data availability statement. Ozone MDA8 data for Southwest PA stations are located at EPA's AQS database: `http://www.epa.gov/ttn/airs/airsaqs/`. GEFS Reforecast v2 dataset with historical weather forecasts is available at `http://www.esrl.noaa.gov/psd/forecasts/reforecast2/`. GFS and NAM MOS data are available at `https://vlab.noaa.gov/web/mdl/mos-archived-data`. Station-specific meteorological data can be found at `https://www.ncei.noaa.gov/products/land-based-station/automated-surface-weather-observing-systems`. NAQFC output is located at the following location `https://www.ncei.noaa.gov/thredds/model/ndgd.html`.

References

Alves, J.-H. G. M., and Coauthors, 2013: The ncep–fnmoc combined wave ensemble product: Expanding benefits of interagency probabilistic forecasts to the oceanic environment. *Bulletin of the American Meteorological Society*, **94 (12)**, 1893 – 1905, https://doi.org/10.1175/BAMS-D-12-00032.1, URL https://journals.ametsoc.org/view/journals/bams/94/12/bams-d-12-00032.1.xml.

Athira, V., P. Geetha, R. Vinayakumar, and K. Soman, 2018: Deepairnet: Applying recurrent networks for air quality prediction. *Procedia computer science*, **132**, 1394–1403.

Avnery, S., D. L. Mauzerall, J. Liu, and L. W. Horowitz, 2011: Global crop yield reductions due to surface ozone exposure: 1. year 2000 crop production losses and economic damage. *Atmospheric Environment*, **45 (13)**, 2284–2296, https://doi.org/https://doi.org/10.1016/j.atmosenv.2010.11.045, URL https://www.sciencedirect.com/science/article/pii/S1352231010010137.

Balashov, N. V., A. M. Thompson, and G. S. Young, 2017: Probabilistic forecasting of surface ozone with a novel statistical approach. *Journal of Applied Meteorology and Climatology*, **56 (2)**, 297 – 316, https://doi.org/10.1175/JAMC-D-16-0110.1, URL https://journals.ametsoc.org/view/journals/apme/56/2/jamc-d-16-0110.1.xml.

Beckett, W. S., 1991: Ozone, air pollution, and respiratory health. *Yale J Biol Med*, **64 (2)**, 167–175.

Bröcker, J., and L. A. Smith, 2007: Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, **22 (2)**, 382 – 388, https://doi.org/10.1175/WAF966.1, URL https://journals.ametsoc.org/view/journals/wefo/22/2/waf966_1.xml.

Campbell, P. C., and Coauthors, 2022: Development and evaluation of an advanced national air quality forecasting capability using the noaa global forecast system version 16. *Geoscientific Model Development*, **15 (8)**, 3281–3313, https://doi.org/10.5194/gmd-15-3281-2022, URL https://gmd.copernicus.org/articles/15/3281/2022/.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the national meteorological center's numerical weather prediction system. *Weather and Forecasting*, **4 (3)**, 401–412.

Chai, T., and Coauthors, 2013: Evaluation of the united states national air quality forecast capability experimental real-time predictions in 2010 using air quality system ozone and no_2 measurements. *Geoscientific Model Development*, **6 (5)**, 1831–1850, https://doi.org/10.5194/gmd-6-1831-2013, URL https://gmd.copernicus.org/articles/6/1831/2013/.

Dabberdt, W. F., and Coauthors, 2004: Meteorological research needs for improved air quality forecasting: Report of the 11th prospectus development team of the u.s. weather research program*. *Bulletin of the American Meteorological Society*, **85 (4)**, 563 – 586, https://doi.org/10.1175/BAMS-85-4-563, URL https://journals.ametsoc.org/view/journals/bams/85/4/bams-85-4-563.xml.

Demeritt, D., S. Nobert, H. Cloke, and F. Pappenberger, 2010: Challenges in communicating and using ensembles in operational flood forecasting. *Meteorological Applications*, **17 (2)**, 209–222, https://doi.org/https://doi.org/10.1002/met.194, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.194, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.194.

Department of Environmental Protection, 2021: Welcome to the pennsylvania air quality partnership. [Online; accessed 01-May-2023], https://www.ahs.dep.pa.gov/aqpartnersweb/.

Djalalova, I., L. Delle Monache, and J. Wilczak, 2015: Pm2.5 analog forecast and kalman filter post-processing for the community multiscale air quality (cmaq) model. *Atmospheric Environment*, **108**, 76–87, https://doi.org/https://doi.org/10.1016/j.atmosenv.2015.02.021, URL https://www.sciencedirect.com/science/article/pii/S1352231015001405.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, B. Yang, and S. Benjamin, 2015: Regional ensemble forecast systems at NCEP. *Preprints, 27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction, Chicago, IL, Amer. Meteor. Soc. A*, Vol. 2.

Eder, B., D. Kang, R. Mathur, J. Pleim, S. Yu, T. Otte, and G. Pouliot, 2009: A performance evaluation of the national air quality forecast capability for the summer of 2007. *Atmospheric Environment*, **43 (14)**, 2312–2320, https://doi.org/https://doi.org/10.1016/j.atmosenv.2009.01.033, URL https://www.sciencedirect.com/science/article/pii/S1352231009000624.

Eder, B., D. Kang, R. Mathur, S. Yu, and K. Schere, 2006: An operational evaluation of the eta–cmaq air quality forecast model. *Atmospheric Environment*, **40 (26)**,

4894–4905, https://doi.org/https://doi.org/10.1016/j.atmosenv.2005.12.062, URL https://www.sciencedirect.com/science/article/pii/S1352231006000938, special issue on Model Evaluation: Evaluation of Urban and Regional Eulerian Air Quality Models.

EPA, 2015: Naaqs table. URL https://www.epa.gov/criteria-air-pollutants/naaqs-table.

Fanfarillo, A., B. Roozitalab, W. Hu, and G. Cervone, 2019: Probabilistic forecasting using deep generative models. *CoRR*, **abs/1909.11865**, URL http://arxiv.org/abs/1909.11865, 1909.11865.

Finley, J. P., 1884: Tornado predictions. *American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches of Study (1884-1896)*, **1 (3)**, 85.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Monthly Weather Review*, **120 (2)**, 361 – 370, https://doi.org/10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/120/2/1520-0493_1992_120_0361_essfcf_2_0_co_2.xml.

Garaud, D., and V. Mallet, 2011: Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality. *Journal of Geophysical Research*, **116 (D19304)**, https://doi.org/10.1029/2011JD015780, URL https://hal.inria.fr/hal-00655771.

Garner, G., and A. Thompson, 2013: Ensemble statistical post-processing of the national air quality forecast capability: Enhancing ozone forecasts in baltimore, maryland. *Atmospheric Environment*, **81**, 517–522, https://doi.org/10.1016/j.atmosenv.2013.09.020.

Gilliam, R. C., C. Hogrefe, J. M. Godowitch, S. Napelenok, R. Mathur, and S. T. Rao, 2015: Impact of inherent meteorology uncertainty on air quality model predictions. *Journal of Geophysical Research: Atmospheres*, **120 (23)**, 12,259–12,280, https://doi.org/https://doi.org/10.1002/2015JD023674, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JD023674, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015JD023674.

Hogan, R. J., C. A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson, 2010: Equitability revisited: Why the "equitable threat score" is not equitable. *Weather and Forecasting*, **25 (2)**, 710 – 726, https://doi.org/10.1175/2009WAF2222350.1, URL https://journals.ametsoc.org/view/journals/wefo/25/2/2009waf2222350_1.xml.

Hogrefe, C., and S. T. Rao, 2001: Demonstrating attainment of the air quality standards: integration of observations and model predictions into the probabilistic framework. *J Air Waste Manag Assoc*, **51 (7)**, 1060–1072.

Joslyn, S., L. Nadav-Greenberg, and R. M. Nichols, 2009: Probability of precipitation: Assessment and enhancement of end-user understanding. *Bulletin of the American Meteorological Society*, **90 (2)**, 185–193.

Kamińska, J. A., 2018: Probabilistic forecasting of nitrogen dioxide concentrations at an urban road intersection. *Sustainability*, **10 (11)**, https://doi.org/10.3390/su10114213, URL https://www.mdpi.com/2071-1050/10/11/4213.

Kang, D., B. K. Eder, A. F. Stein, G. A. Grell, S. E. Peckham, and J. McHenry, 2005: The new england air quality forecasting pilot program: Development of an evaluation protocol and performance benchmark. *Journal of the Air & Waste Management Association*, **55 (12)**, 1782–1796, https://doi.org/10.1080/10473289.2005.10464775, URL https://doi.org/10.1080/10473289.2005.10464775, https://doi.org/10.1080/10473289.2005.10464775.

Kleinman, L. I., W. F. Ryan, P. H. Daum, S. R. Springston, Y.-N. Lee, L. J. Nunnermacker, and J. Weinstein-Lloyd, 2004: An ozone episode in the Philadelphia metropolitan area. *Journal of Geophysical Research (Atmospheres)*, **109 (D20)**, D20302, https://doi.org/10.1029/2004JD004563.

Kohonen, T., 2013: Essentials of the self-organizing map. *Neural networks*, **37**, 52–65.

Lee, D. S., M. R. Holland, and N. Falla, 1996: The potential impact of ozone on materials in the U.K. *Atmospheric Environment*, **30 (7)**, 1053–1065, https://doi.org/https://doi.org/10.1016/1352-2310(95)00407-6, URL https://www.sciencedirect.com/science/article/pii/1352231095004076.

Lee, P., and Coauthors, 2017: Naqfc developmental forecast guidance for fine particulate matter (pm2.5). *Weather and Forecasting*, **32 (1)**, 343 – 360, https://doi.org/https://doi.org/10.1175/WAF-D-15-0163.1, URL https://journals.ametsoc.org/view/journals/wefo/32/1/waf-d-15-0163_1.xml.

Lin, M., L. W. Horowitz, R. Payton, A. M. Fiore, and G. Tonnesen, 2017: US surface ozone trends and extremes from 1980 to 2014: quantifying the roles of rising asian emissions, domestic controls, wildfires, and climate. *Atmospheric Chemistry and Physics*, **17 (4)**, 2943–2970, https://doi.org/10.5194/acp-17-2943-2017, URL https://acp.copernicus.org/articles/17/2943/2017/.

Mason, I., 1979: On reducing probability forecasts to yes/no forecasts. *Monthly Weather Review*, **107 (2)**, 207 – 211, https://doi.org/10.1175/1520-0493(1979)107<0207:ORPFTY>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/107/2/1520-0493_1979_107_0207_orpfty_2_0_co_2.xml.

Monache, L. D., and Coauthors, 2008: A kalman-filter bias correction method applied to deterministic, ensemble averaged and probabilistic forecasts of surface ozone. *Tellus B: Chemical and Physical Meteorology*, **60 (2)**, 238–249, https://doi.org/10.1111/j.1600-0889.2007.00332.x, URL https://doi.org/10.1111/j.1600-0889.2007.00332.x, https://doi.org/10.1111/j.1600-0889.2007.00332.x.

National centers for environmental information, 2022: National digital guidance database. [Online; accessed 17-October-2022], https://www.ncei.noaa.gov/thredds/model/ndgd.html.

NOAA Air Resources Laboratory, 2019: Arl advances major upgrade to naqfc emission science. [Online; accessed 28-April-2023], https://www.arl.noaa.gov/news-pubs/news-archive/arl-advances-major-upgrade-to-naqfc-emission-science-2/.

Pinder, R. W., R. C. Gilliam, K. W. Appel, S. L. Napelenok, K. M. Foley, and A. B. Gilliland, 2009: Efficient probabilistic estimates of surface ozone concentration using an ensemble of model configurations and direct sensitivity calculations. *Environ Sci Technol*, **43 (7)**, 2388–2393.

Ramos, M. H., S. J. van Andel, and F. Pappenberger, 2013: Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, **17 (6)**, 2219–2232, https://doi.org/10.5194/hess-17-2219-2013, URL https://hess.copernicus.org/articles/17/2219/2013/.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, **130 (6)**, 1653 – 1660, https://doi.org/10.1175/1520-0493(2002)

130<1653:EPFUIT>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/130/6/1520-0493_2002_130_1653_epfuit_2.0.co_2.xml.

Ryan, W. F., C. A. Piety, and E. D. Luebehusen, 2000: Air quality forecasts in the mid-atlantic region: Current practice and benchmark skill. *Weather and Forecasting*, **15 (1)**, 46 – 60, https://doi.org/10.1175/1520-0434(2000)015<0046:AQFITM>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/15/1/1520-0434_2000_015_0046_aqfitm_2_0_co_2.xml.

Schneider, R. S., A. R. Dean, and H. Brooks, 2009: Estimating potential severe weather societal impacts using probabilistic forecasts issued by the NWS storm prediction center. *23rd Conf. on Weather and Forecasting*.

Stajner, I., P. Davidson, D. Byun, J. McQueen, R. Draxler, P. Dickerson, and J. Meagher, 2012: Us national air quality forecast capability: Expanding coverage to include particulate matter. *Air Pollution Modeling and its Application XXI*, D. G. Steyn, and S. Trini Castelli, Eds., Springer Netherlands, Dordrecht, 379–384.

Vautard, R., and Coauthors, 2006: Is regional air quality model diversity representative of uncertainty for ozone simulation? *Geophysical Research Letters*, **33 (24)**.

Weather prediction center, 2021: WPC probabilistic winter precipitation guidance. [Online; accessed 13-July-2021], https://www.wpc.ncep.noaa.gov/pwpf/wwd_accum_probs.php.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*, Vol. 100. Academic press.