Atmospheric Environment 81 (2013) 517–522

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Ensemble statistical post-processing of the National Air Quality Forecast Capability: Enhancing ozone forecasts in Baltimore, Maryland

Gregory G. Garner [a,*], Anne M. Thompson [b]

[a] The Pennsylvania State University, 503 Walker Building, University Park, PA 16802, USA
[b] NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

HIGHLIGHTS

- We developed an air quality forecast tool for Baltimore, MD.
- The tool adjusts air quality model output for local meteorology.
- The tool consists of bootstrapped regression trees and extreme-value theory.
- The tool provides value for decisions which the model alone could not.
- Applications to particulate matter forecasts are promising.

ARTICLE INFO

Article history:
Received 21 June 2013
Received in revised form
10 September 2013
Accepted 13 September 2013

2010 MSC:
62P12
62G32

Keywords:
Ensemble forecast
Air quality
Baltimore
Regression tree
Extreme value
Bootstrap

ABSTRACT

An ensemble statistical post-processor (ESP) is developed for the National Air Quality Forecast Capability (NAQFC) to address the unique challenges of forecasting surface ozone in Baltimore, MD. Air quality and meteorological data were collected from the eight monitors that constitute the Baltimore forecast region. These data were used to build the ESP using a moving-block bootstrap, regression tree models, and extreme-value theory. The ESP was evaluated using a 10-fold cross-validation to avoid evaluation with the same data used in the development process. Results indicate that the ESP is conditionally biased, likely due to slight overfitting while training the regression tree models. When viewed from the perspective of a decision-maker, the ESP provides a wealth of additional information previously not available through the NAQFC alone. The user is provided the freedom to tailor the forecast to the decision at hand by using decision-specific probability thresholds that define a forecast for an ozone exceedance. Taking advantage of the ESP, the user not only receives an increase in value over the NAQFC, but also receives value for costly decisions that the NAQFC couldn't provide alone.

1. Introduction

Baltimore, MD is ranked the ninth most ozone polluted metropolitan area in the U.S. according to the 2013 State of the Air released by the American Lung Association (http://www.stateoftheair.org/2013), sharing the top spots with cities such as Los Angeles, CA and Houston, TX. Consequently, forecasting ozone is vital to the health and well being of over eight million Baltimore residents. Expert air quality forecasters use a combination of numerical models, statistical models, and empirical rules to provide accurate and timely forecasts. These tools generally provide skillful (Eder et al., 2010) and valuable (Garner and Thompson, 2012) forecasts; however, these tools have limitations that pose a unique forecast challenge.

The Baltimore forecast region is depicted in Fig. 1. The region encompasses counties of Maryland that are located to the north and west of the Chesapeake Bay. Eight ozone monitors, described in Table 1, comprise the monitoring network within the Baltimore

* Corresponding author.
E-mail address: ggg121@psu.edu (G.G. Garner).
URL: http://www.meteo.psu.edu/%7Eggg121

1352-2310
http://dx.doi.org/10.1016/j.atmosenv.2013.09.020

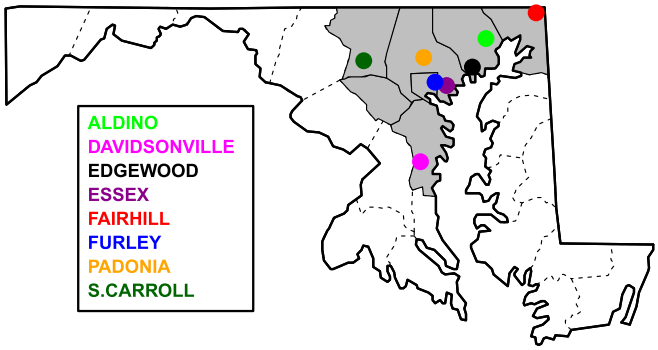

Fig. 1. Map of ozone monitors in the Baltimore, MD forecast region. The forecast region is shaded in gray according to the region definition provided by the Maryland Department of the Environment.

forecast region. Six of the eight monitors lie within 30 km of the bay.

Near-surface ozone is produced through photochemical reactions with nitrogen oxides (NO_x) and volatile organic compounds (VOCs) (Seinfeld and Pandis, 2006); thus seasonal ozone concentrations peak during the late-spring through early-fall months (April–October; Ozone Season) when ample solar radiation is available for the photochemical reactions. Daily ozone concentrations, however, are strongly driven by the local meteorology. Cloud-free skies maximize the actinic flux available to the photochemical reactions while light winds promote stagnation and aggregation of ozone and its precursors. These meteorological features also create an ideal environment through which a bay breeze may form within the Baltimore forecast region (Banta et al., 2005). The temperature gradient between the warm solar-heated land and the cool water creates a thermal circulation which can concentrate ozone and its precursors along the airmass boundary (Stauffer et al., 2012; Stauffer and Thompson, 2013). The current suite of regional operational numerical models are run at resolutions that are too coarse to forecast the onset and location of bay breeze events to the degree needed for assessing their impacts on local air quality (Banta et al., 2005). This includes the National Air Quality Forecast Capability (NAQFC), the current national air quality model produced by the National Oceanic and Atmospheric Administration (NOAA) and the Environmental Protection Agency (EPA) (Janjic, 2003; Byun and Schere, 2006; Garner et al., 2013), with an operational horizontal resolution of 12 km. Loughner et al. (2011) found that a horizontal resolution of 4.5 km or finer in numerical meteorological models produced discernible simulations of the bay breeze in the Baltimore region. Without properly resolving the bay breeze, the NAQFC will not be able to properly handle ozone predictions along coastal boundaries such as those in the Baltimore forecast region.

Attaining uncertainty about the forecast from a numerical model is difficult. Common practice is to run an ensemble of numerical models, each with slightly perturbed initial conditions, boundary conditions, and/or parameterizations such as the NOAA

Table 1
Baltimore, MD air quality monitor locations.

Site Name	FIPS Code	Lat [deg N]	Lon [deg E]	Elev. [m]
Aldino	24-025-9001	39.563	−76.204	127.7
Davidsonville	24-003-0014	38.903	−76.653	44.0
Edgewood	24-025-1001	39.410	−76.297	8.5
Essex	24-005-3001	39.311	−76.474	12.8
Fairhill	24-015-0003	39.701	−75.860	117.7
Furley	24-510-0054	39.329	−76.553	49.0
Padonia	24-005-1007	39.461	−76.631	119.5
South Carrol	24-013-0001	39.444	−77.042	226.0

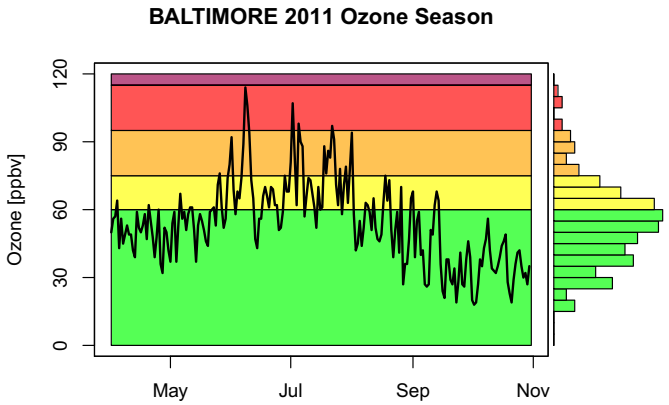

Fig. 2. Time-series of the 2011 daily maximum 8-hr average ozone in Baltimore, MD. Ozone values are the daily maximum 8-hr averages among all the sites listed in Table 1. The ozone-season is defined as 01 April through 31 October. The histogram along the right margin is positively skewed suggesting that statistical models built on assumptions of normality using these data may result in underforecasting the ozone exceedance events. The background is shaded according to the air quality index.

Short Range Ensemble Forecast (SREF; http://www.spc.noaa.gov/exper/sref/fplumes/). From the suite of models used in the ensemble, one can determine a consensus prediction and spread from which uncertainties are derived. In order to reduce the computational burden, often times the members within the ensemble are run at a reduced resolution. This process is not feasible for the NAQFC for reasons described earlier. Alternatively, air quality forecasters are using multiple numerical models from different sources to create a "poor-man's" ensemble (Djalalova et al., 2010). Though this method was shown to improve upon the forecast from any single ensemble member, the ensemble would rely heavily on the individual model providers to continue producing the forecasts. If a single model provider decides to terminate their model, due to lack of funding for example, then the entire ensemble suffers.

In addition to numerical models, air quality forecasters often use statistical models derived from local data. These statistical models range from simple regression models to complex artificial neural networks (Thompson et al., 2001; Al-Alawi et al., 2008; Pires and Martins, 2011). The skewed distribution of ozone, as evidenced in Fig. 2, means that many of the standard statistical approaches may not be valid and that methods that incorporate extreme-value theory are preferred (Thompson et al., 2001).

The goal of this article is to develop a tool that can address these problems while providing the best value of information to the end user. The product must adjust appropriately for various local meteorological regimes and provide uncertainty information about the forecast while avoiding the pitfalls of forecasting extreme values. An ensemble statistical post-processor (ESP) for the NAQFC is a logical choice for such a product.

2. Data and methods

Data were collected for eight air quality monitoring locations in the Baltimore, MD forecast region shown in Fig. 1 and described in Table 1. Hourly ozone observations from the 2005–2011 ozone seasons (April–October) were collected from the EPA Technology Transfer Network (TTN) Air Quality System (AQS), a quality-controlled national database of atmospheric particle and trace-gas measurements (http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm). The National Ambient Air Quality Standard (NAAQS) for ozone is calculated using forward-running 8-h average concentrations, so the hourly ozone data were averaged as

Table 2
Meteorological variables used in the development of the ESP product.

Variable	Time/type	Units
Temperature	Max, Min, 1800 UTC	K
Dewpoint Temperature	Max, Min, 1800 UTC	K
Sea-level Pressure	Max, Min	hPa
Relative Humidity	Max, Min	%
Sky Cover	1200 UTC, 1800 UTC	%
U-component Wind	1200 UTC, 1800 UTC	ms^{-1}
V-component Wind	1200 UTC, 1800 UTC	ms^{-1}
Precipitation	24-h Total	mm
Bay Breeze Index	1800 UTC	–

such. For any given hour, the ozone concentration for that hour is averaged with the subsequent seven hours of ozone data. This average represents the 8-h average for that particular hour. This averaging is performed for each hour in the data set. The daily maximum 8-h averages are used for the product development and validation. An ozone exceedance is defined in the NAAQS as a day with a maximum 8-h average ozone concentration greater than 75 ppbv.

The daily 1200 UTC NAQFC model output was collected from the NOAA National Operational Model Archive and Distribution System for the same dates as the hourly ozone observations. The NAQFC model output includes the 8-h running average ozone, so no additional averaging is necessary. The maximum 8-h average ozone forecasted for the day following the initial model run date is used.

Historical meteorological observations collocated with the ozone monitors were collected from the National Climatic Data Center. In the event that meteorological information is not available for an ozone monitor, meteorological data from surrounding sites were interpolated to the monitor using a kriging algorithm (Ribeiro and Diggle, 2001). The meteorological data set is listed in Table 2. The bay breeze index is a derived quantity (Sikora et al., 2010) that represents the degree to which the atmosphere favors bay breeze formation. The index is calculated relative to meteorological data from the buoy station TPLM2 at Thomas Point, MD. Including such an index will help the model development process sort out days with possible bay breeze events.

2.1. ESP development

The ESP development mimics that of a perfect prog system (Klein et al., 1959; Wilks, 2011). Observational data are used to train the ESP. Forecasted quantities of the variables used in development are then used in the ESP to predict the future ozone concentrations. This method is preferred over model output statistics (Glahn and Lowry, 1972) when forecasting air quality due to the young age and rapid development of the NAQFC. As the NAQFC matures, the increase in forecast skill would translate over into the ESP.

This development process is applied to each monitor separately. The meteorological variables in Table 2 along with the NAQFC data are used as independent variables to predict the dependent variable ozone. The ozone concentration from the previous day is also included as an independent variable in order to take advantage of any serial correlation in the ozone data. These data are resampled 100 times with replacement using a moving-block bootstrap algorithm to produce 100 bootstrap subsamples of equal length of the original data set (Efron, 1979; Efron and Tibshirani, 1993; Wilks, 2011).

Each bootstrap subsample is fit to a unique regression tree model (Breiman, 1984). The purpose of a regression tree model is to recursively split the ozone into homogeneous groups called nodes using the independent data. Splits and terminal nodes are typically scored with a metric based on the standard deviation of the groups, but in order to better predict an ozone exceedance, the f-measure is used instead (Torgo and Ribeiro, 2003; Ribeiro and Torgo, 2006). The f-measure

$$F = \frac{\left(\beta^2 + 1\right) \cdot precision \cdot recall}{\beta^2 \cdot precision \cdot recall} \tag{1}$$

is rooted in extreme-value theory and used to predict outliers of a data set by accounting not only for the homogeneity of the group, but also the ability of the group to recall the extreme values. The variable *precision* is 1 − NMSE or a function of the normalized mean square error of the predictions for extreme values. The variable *recall* is the ratio of predicted extreme values to the number of extreme values in the data set. β is a parameter that adjusts the relative importance of *precision* to *recall*. By defining an extreme value of ozone as the NAAQS of 75 ppbv and using the f-measure as a split function in these regression trees, the terminal nodes will be tailored to ozone exceedances. The path to the terminal nodes will represent local meteorological regimes conducive to producing ozone exceedances. The terminal nodes contain homogeneous clusters of ozone and the independent data describing it. Multivariate linear regression models are then fit to the data in the terminal nodes.

The resulting 100 regression tree models constitute the ESP for the given monitor. Three parameters are used in the development process. First, the minimum number of data instances for a given node was set to 30. This ensures enough data in a terminal node from which a regression model can be fit. Second and third, the β parameter in the f-measure and the node-termination threshold of the f-measure were set to 0.8. These values were selected to give an advantage to the *recall* term in the f-measure and allow the tree to grow to a reasonable number of terminal nodes. The ESP product can be used operationally to adjust the ozone concentration from the NAQFC using forecasted values of the meteorological variables.

2.2. Cross-validation

The ESP product was evaluated using a 10-fold cross-validation scheme (Picard and Cook, 1984; Wilks, 2011). The full data set is split into 10 groups each containing 10% of the original data set. Nine of the groups are used in the development process described in Section 2.1 while the tenth group is reserved for evaluation. This process is repeated until each of the 10 groups is used as a reserved evaluation data set. Cross-validation ensures that the product is never evaluated with the same data used to build the product, resulting in evaluation metrics that closely represent skill in true operational forecasts.

3. Results

Probabilities are derived from the ensemble predictions. The number of predictions above the ozone standard of 75 ppbv is divided by the number of available ensemble predictions. This results in the forecasted probability of the monitor exceeding the ozone standard for the day. Baltimore regional probability forecasts were calculated from the individual monitor probability forecasts through a recursive application of the additive law of probability for non-mutually exclusive events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2}$$

This expression produces the forecasted probability of any given site within the Baltimore forecast region exceeding the ozone

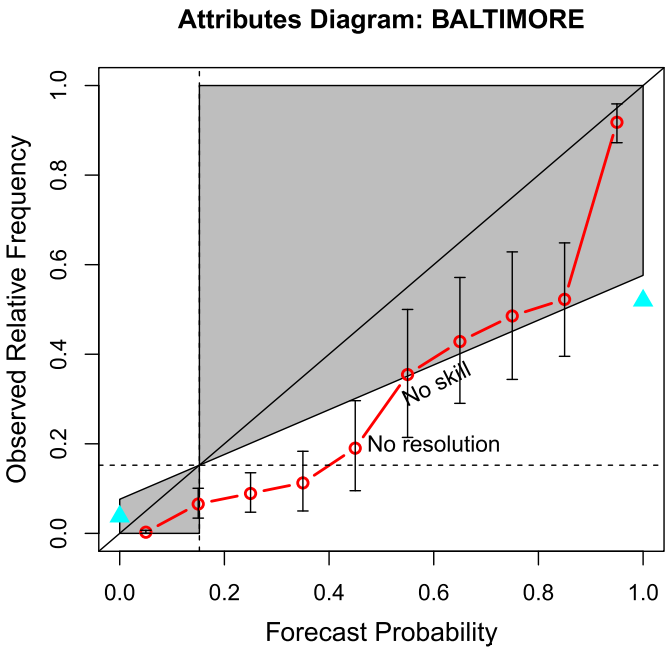

Fig. 3. Attributes diagram for the ESP product for the Baltimore, MD forecast region. The observed relative frequency of an exceedance event is plotted as a function of the forecasted probability based on the ESP product. An ideal diagram would follow the 1:1 line indicating that the forecasted probability perfectly matches the observed frequency of exceedances given the forecast. The error bars represent the 95% confidence intervals about the mean observed relative frequency for a given forecast probability derived empirically from 10,000 bootstrapped subsamples. The forecast probabilities are binned into 10% bins to provide enough sample points from which to derive confidence intervals as well as facilitate interpretation. The triangles are the NAQFC forecasts converted into a binary forecasts for ozone exceedances using the NAAQS threshold of 75 ppbv.

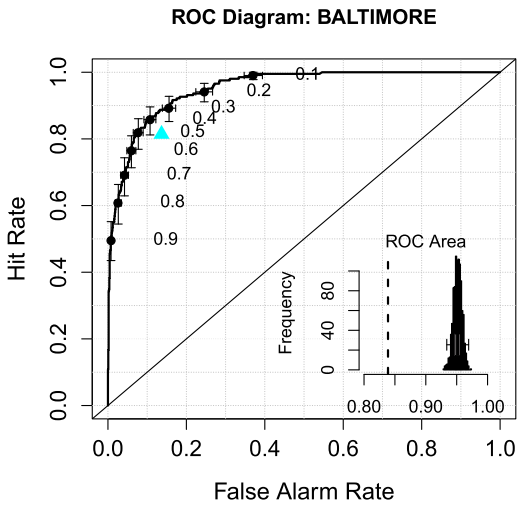

Fig. 4. Relative operating characteristic (ROC) diagram for the ESP product for the Baltimore, MD forecast region. The hit rate is plotted as a function of the false alarm rate for a series of forecast probability thresholds which define a forecasted exceedance. The dots represent the different forecast probability thresholds used to convert the probabilistic forecast into a binary forecast. The error bars are the 95% confidence interval about the mean hit rate (vertical) and false alarm rate (horizontal) derived empirically from 10,000 bootstrap subsamples. The triangle is the NAQFC forecast. The histogram inset describes the distribution of the area under the ROC curve based on the bootstrap subsamples used in deriving the confidence intervals. The vertical dashed line in the inset represents the area of the ROC curve based on the NAQFC forecast.

standard. Results of the ESP cross-validation for the Baltimore forecast region are shared here. Please refer to the provided Supplementary material for the ESP cross-validation results of each monitor individually.

The attributes diagram in Fig. 3 describes the full joint distribution of the ESP forecasts and observed ozone in the Baltimore region (Hsu and Murphy, 1986). The observed relative frequency of ozone exceedances is plotted as a function of the ESP forecasted probability. The points in this diagram would lie on the 1:1 line when using a perfect ensemble forecast system indicating that, for example, ozone exceeds 30% of the time when the forecasted probability is 30%. A point falling on the "No resolution" line indicates that the associated subset of forecasts are unable to discern events that are different from the climatological probability of the event. The "No skill" line is half-way between the ideal 1:1 line and the "No resolution" line and defines the region where forecasts produce positive skill (shaded) versus negative skill (non-shaded). The points lie below the 1:1 line at 95% significance with the exception of the highest forecast probabilities. This diagram indicates that the ESP tends to overpredict the frequency of ozone exceedances. Observed frequency of ozone exceedances associated with forecast probability between 0.3 and 0.5 are not significantly different from the climatological frequency. Forecast probabilities between 0.5 and 0.9 on average produce marginal positive skill for the ESP, though not statistically significant. Most of the skill of the ESP lies in the lowest (0–0.2) and highest (0.9–1) forecast probabilities. This conditional bias is expected because the ESP was developed with the intent of forecasting ozone exceedances, thus sacrificing accurate predictions for middle-frequency events. Some of the conditional bias may also be attributed to overfitting during the regression tree training process. The triangles are the NAQFC forecasts converted into binary forecasts of either exceedance (NAQFC $>$ 75 ppbv or probability $=$ 1) or non-exceedance (NAQFC $\leq$ 75 ppbv or probability $=$ 0) and are provided for reference. When the NAQFC forecasts an exceedance event, the event occurs just above 50% of the time. When the NAQCF forecasts a non-exceedance event, the event occurs approximately 5% of the time.

The relative operating characteristic (ROC) curve (Swets, 1979; Mason, 1982; Wilks, 2011) shown in Fig. 4 provides an analysis of forecast performance from the perspective of a decision maker. The ESP hit rate is plotted as a function of the ESP false alarm rate. The probabilistic forecasts from the ESP are converted into binary predictions using a threshold of probability. For example, a threshold of 0.5 means that any forecast with a probability of exceedance greater than 0.5 would be a forecast for an exceedance while a forecast probability less than 0.5 would be a forecast for a non-exceedance. The hit rate is the proportion of correctly predicted exceedances to the number of observed exceedances while the false alarm rate is the proportion of incorrectly forecasted exceedances to the total number of non-exceedances. Points on the ROC curve are associated with various thresholds of forecast probability that determine a forecast for exceedance. The ideal curve would create a right-angle in the upper-left corner of the plot suggesting there would be a probability threshold that produces a perfect hit rate while never producing a false alarm. Interpretation of this plot depends on the needs of the user. A user whose decision is sensitive to the false alarm rate of the forecast system may use this plot to identify a forecast probability threshold that maximizes the hit rate while remaining below an acceptable false alarm rate. For example, such a user would be able to achieve a 92% hit rate while remaining below the 20% false alarm rate using the ESP and a probability threshold of 35%. Compared to the NAQFC, the ESP produces a 7.5% increase hit rate at the same false alarm rate and a 4.2% decrease in false alarm rate at the same hit rate. The distribution of ROC area (area under the ROC curve) for the ESP is tightly centered around a

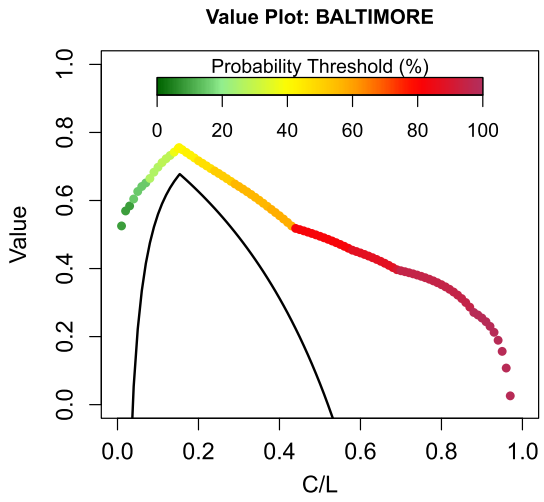

Fig. 5. Value curve for the ESP product for the Baltimore, MD forecast region. Color shading represents the probability threshold used to get the maximum value for the given cost-loss ratio. The solid black line is the value curve for the NAQFC.

mean of 0.95 and is significantly different from the ROC area associated with the NAQFC (vertical dashed line).

The goal of developing the ESP was to address the difficulties in forecasting ozone in Baltimore while providing increased value of information. Fig. 5 depicts the increase in value of the air quality forecasts when using the ESP relative to the NAQFC alone. Value is calculated using a static cost-loss ratio model (Thompson, 1952; Richardson, 2000; Garner and Thompson, 2012; Garner et al., 2013). Value is the percent savings in expenditure over climatological expenditure relative to a perfect forecast system. The cost-loss ratio (C/L) is defined by the cost (C) to insure against loss (L) of an ozone exceedance. For a given loss, C/L can be interpreted as a suite of possible decisions relevant to the user. For example, the maximum value of the ESP is 0.75 at a C/L of 0.18 using a probability threshold of 44%. This means that if the user makes a decision based on the ESP to implement a program that costs \$18,000 to prevent a loss of \$100,000, the user can expect to save \$11,070 per event $[18\% \times (\$100,000 - \$18,000) \times 75\% = \$11,070]$. The benefit of an ensemble forecast system is that the user can maximize the value of the forecasts for multiple decisions by choosing multiple probability thresholds that define a forecasted exceedance. Say that the same user from before would like to attain the maximum value for their decision costing \$60,000. This user would use a probability threshold of 84% for this decision as opposed to the 44% threshold used earlier. The ESP not only provides an increase in value of information over the NAQFC alone (black curve in Fig. 5), but the ensemble approach also provides value at C/L far exceeding what the NAQFC could provide. This means that the NAQFC, after application of the ESP, can not only provide more value for decisions currently covered by the NAQFC, but also value for high-cost decisions currently not covered by the NAQFC.

4. Summary and discussion

The ESP was developed to address the challenges in forecasting air quality in Baltimore, MD. These challenges include local meteorological phenomena unresolvable by the current numerical models, uncertainty about air quality forecasts, and pitfalls in statistical assumptions when fitting standard statistical models.

Ozone and meteorological data were collected from eight ozone monitors that represent the Baltimore forecast region. These data were used to develop 100 regression trees for each monitor using a moving-block bootstrap algorithm. The 100 regression trees constitute the ESP for the given monitor. Each regression tree is fitted using the f-measure (Eq. (1)) for node splitting and evaluation. The result is regression trees with meteorology-dependent paths leading to nodes tailored to predicting ozone exceedances. The ESP was evaluated with a 10-fold cross-validation designed to mimic an operational forecasting scenario.

Results indicate that the ESP exhibits conditional bias. This conditional bias was expected due to the measures taken to achieve the goal of the product. In addition, some overfitting of the regression tree models may have contributed to the conditional bias. The ESP tends to overpredict the relative frequency of ozone exceedances between forecast probabilities of 0.2–0.9. The skill of the ESP is shown at the extremes for forecast probabilities. Individual monitors behave better in this regard when compared to the region as a whole.

From the perspective of a decision maker, the ESP provides significantly more information than the NAQFC alone. The ESP allows the user to choose probability thresholds that fit the criteria of their decision, such as sensitivity to false alarms or minimum achievable hit rate. The ESP was also shown to not only improve the value of the NAQFC, but also provide significant value at decisions not attainable with the NAQFC alone.

The ESP can be a valuable tool to an air quality forecaster. This article describes the ESP development for forecasting ozone in Baltimore, but these methods can be easily adapted and applied to many other forecast challenges.

Acknowledgments

The author would like to acknowledge Dan Salkovitz from the Virginia Department of Environmental Quality and Laura Warren from the Maryland Department of the Environment for their useful comments and feedback on the statistical guidance product. This research was supported by a STAR fellowship (FP-91729901-1) to GGG awarded by the U.S. Environmental Protection Agency (EPA). It has not been formally reviewed by the EPA. The views expressed in this manuscript are solely those of GGG and co-authors. The EPA does not endorse any products or commercial services mentioned in this manuscript. Additional funding for this research was provided by grants to the Pennsylvania State University from NASA DISCOVER-AQ (NNX10AR39G), the NASA Air Quality Applied Sciences Team (NNX11AQ44G).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2013.09.020.

References

Al-Alawi, S.M., Abdul-Wahab, S.A., Bakheit, C.S., 2008. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. Environ. Model. Softw. 23, 396–403. http://dx.doi.org/10.1016/j.envsoft.2006.08.007.

Banta, R.M., Senff, C.J., Neilson-Gammon, J., Darby, L.S., Ryerson, T.B., Alvarez, R.J., Sandberg, S.P., Williams, E.J., Trainer, M., 2005. A bad air day in houston. Bull. Am. Meteorol. Soc. 86, 657–669. http://dx.doi.org/10.1175/BAMS-86-5-657.

Breiman, L., 1984. Classification and Regression Trees. In: Wadsworth Statistics/probability Series. Wadsworth International Group.

Byun, D., Schere, K.L., 2006. Review of the governing equations, computational algorithms, and other components of the models-3 Community Multiscale Air Quality (CMAQ) modeling system. Appl. Mech. Rev. 59, 51–77.

Djalalova, I., Wilczak, J., McKeen, S., Grell, G., Peckham, S., Pagowski, M., DelleMonache, L., McQueen, J., Tang, Y., Lee, P., McHenry, J., Gong, W., Bouchet, V., Mathur, R., 2010. Ensemble and bias-correction techniques for air quality model forecasts of surface O3 and PM2.5 during the Texaqs-ii experiment of 2006. Atmos. Environ. 44, 455–467. http://dx.doi.org/10.1016/j.atmosenv.2009.11.007.

Eder, B., Kang, D., Rao, S.T., Mathur, R., Yu, S., Otte, T., Schere, K., Wayland, R., Jackson, S., Davidson, P., McQueen, J., Bridgers, G., 2010. Using national air quality forecast guidance to develop local air quality index forecasts. Bull. Am. Meteorol. Soc. 91, 313–326. http://dx.doi.org/10.1175/2009BAMS2734.1.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall.

Garner, G.G., Thompson, A.M., 2012. The value of air quality forecasting in the mid-Atlantic region. Wea. Clim. Soc. 4, 69–79. http://dx.doi.org/10.1175/WCAS-D-10-05010.1.

Garner, G.G., Thompson, A.M., Lee, P., Martins, D.K., 2013. Evaluation of naqfc model performance in forecasting surface ozone during the 2011 discover-aq campaign. J. Atmos. Chem., 1–19. http://dx.doi.org/10.1007/s10874-013-9251-z.

Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics *mos* in objective weather forecasting. J. Appl. Meteorol. 11, 1203–1211.

Hsu, W.R., Murphy, A.H., 1986. The attributes diagram a geometrical framework for assessing the quality of probability forecasts. Int. J. Forecast. 2, 285–293. http://dx.doi.org/10.1016/0169-2070(86)90048-8.

Janjic, Z.I., 2003. A nonhydrostatic model based on a new approach. Meteorl. Atmos. Phys. 82, 271–285.

Klein, W.H., Lewis, B.M., Enger, I., 1959. Objective prediction of five-day mean temperatures during winter. J. Meteorol. 16, 672–682.

Loughner, C.P., Allen, D.J., Pickering, K.E., Zhang, D.L., Shou, Y.X., Dickerson, R.R., 2011. Impact of fair-weather cumulus clouds and the Chesapeake Bay breeze on pollutant transport and transformation. Atmos. Environ. 45, 4060–4072. http://dx.doi.org/10.1016/j.atmosenv.2011.04.003.

Mason, I., 1982. A model for assessment of weather forecasts. Aust. Meteorol. Mag. 30, 291–303.

Picard, R.R., Cook, R.D., 1984. Cross-validation of regression models. J. Am. Stat. Assoc. 79, 575–583.

Pires, J., Martins, F., 2011. Correction methods for statistical models in tropospheric ozone forecasting. Atmos. Environ. 45, 2413–2417. http://dx.doi.org/10.1016/j.atmosenv.2011.02.011.

Ribeiro, R., Torgo, L., 2006. Rule-based prediction of rare extreme values. In: Todorovski, L., Lavrac, N., Jantke, K. (Eds.), Discovery Science, Lecture Notes in Computer Science, vol. 4265. Springer, Berlin/Heidelberg, pp. 219–230.

Ribeiro Jr., P.J., Diggle, P.J., 2001. geoR: a package for geostatistical analysis. R. News 1, 14–18.

Richardson, D.S., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. Q. J. R. Meteorol. Soc. 126, 649–667.

Seinfeld, J.H., Pandis, S.N., 2006. Chemistry of the troposphere. In: Atmospheric Chemistry and Physics: from Air Pollution to Climate Change, second ed. John Wiley and Sons Inc., pp. 205–283.

Sikora, T.D., Young, G.S., Bettwy, M.J., 2010. Analysis of the western shore Chesapeake Bay bay-breeze. Natl. Weather Dig. 34, 55–65.

Stauffer, R.M., Thompson, A.M., 2013. Bay breeze climatology at two sites along the Chesapeake Bay from 1986–2010: implications for surface ozone. J. Atmos. Chem., 1–18. http://dx.doi.org/10.1007/s10874-013-9260-y.

Stauffer, R.M., Thompson, A.M., Martins, D.K., Clark, R.D., Goldberg, D.L., Loughner, C.P., Delgado, R., Dickerson, R.R., Stehr, J.W., Tzortziou, M.A., 2012. Bay breeze influence on surface ozone at Edgewood, MD during July 2011. J. Atmos. Chem., 1–19. http://dx.doi.org/10.1007/s10874-012-9241-6.

Swets, J.A., 1979. ROC analysis applied to the evaluation of medical imaging techniques. Investig. Radiol. 14, 109–121.

Thompson, J.C., 1952. On the operational deficiencies in categorical weather forecasts. Bull. Am. Meteorol. Soc. 33, 223–226.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P., Sampson, P.D., 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmos. Environ. 35, 617–630. http://dx.doi.org/10.1016/S1352-2310(00)00261-2.

Torgo, L., Ribeiro, R., 2003. Predicting outliers. In: Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H. (Eds.), Knowledge Discovery in Databases: PKDD 2003, Lecture Notes in Computer Science, vol. 2838. Springer, Berlin/Heidelberg, pp. 447–458.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences, third ed. Elsevier.