REACHING SCIENTIFIC CONSENSUS THROUGH A COMPETITION

BY VALLIAPPA LAKSHMANAN, KIMBERLY L. ELMORE, AND MICHAEL B. RICHMAN

WHY A COMPETITION? In a manner similar to the other Science and Technology Advisory Committees (STACs), the AMS's STAC for Artificial Intelligence (AI) conducts specialty scientific conferences. We noticed that for the most part, the AI conferences consisted of researchers who were not fully engaged in each others' presentations. To some extent, this problem of people talking but not listening is unique to AI in meteorology, but we suspect that the dynamics that make this pervasive in AI also exist in other specialties.

AI consists of techniques that employ computers to find solutions to problems that would otherwise have to be performed at a considerable outlay of time or effort by humans. AI borrows from applied statistics, signal processing, and computer science to solve problems through automation.

In meteorology, AI has been used to address issues such as estimating rainfall amounts, nowcasting lightning, predicting convective initiation, diagnosing tornado probability, controlling radar data quality, and approximating computationally expensive models.

Speakers at AI conferences typically expound on the problem at hand and the approach they followed to solve it. Unfortunately, the researchers who would be knowledgeable about the problem being solved would more likely be at the hydrology, lightning, GOES, or nowcasting conferences. The audience at the AI conference tends to consist of researchers interested in AI. Accordingly, the specifics of the problem that motivated the particular solution would be outside the expertise of the audience. Yet, there is no way to successfully exchange scientific knowledge between researchers in AI without understanding the problem at hand, mainly because the selection of AI method (it was thought) depended heavily on the problem being solved.

Based on this supposition, we decided to have one session at our annual AI meetings be a ”competition.” Someone would put up and explain a dataset. Then, a variety of researchers would apply different techniques to the AI dataset. At the conference, the dataset would be described in detail, and every speaker would recount the characteristics of the problem that motivated the methodology that was used.

In order to pique interest and increase the number of techniques applied to address the problem, we cast it as a competition. Entries would be ranked on a predetermined measure of skill, with a special prize for the most skilled student entry. Private-sector companies with an interest in the problem being addressed—Weather Decision Technologies (WDT) in the first year and WSI Corporation in the second—donated money for the prizes.

ANYTHING GOES. In the first year (2007–08) of the competition, the challenge was to identify the type of storm—supercell, convective line, pulse storm, or unorganized—based on attributes derived from radar data by a storm-tracking algorithm. The training data were supplied by Guillot et al. (2008) and consisted of the storm type of various storms as identified by a human researcher. Attributes of these storms were extracted by the storm-tracking method of Lakshmanan and Smith described in a 2009 article in the *Journal of Atmospheric and Oceanic Technology*. These attributes, as well as the human categorization of the storms by type, were provided to the contestants. The contestants used this training dataset to create their AI models. A separate test dataset, consisting of a similar variety of storms, was later provided to the contestants but without the human storm-type cat-

AFFILIATIONS: LAKSHMANAN AND ELMORE—Cooperative Institute of Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma, and National Oceanic and Atmospheric Administration/National Severe Storms Laboratory, Norman, Oklahoma; RICHMAN—School of Meteorology, University of Oklahoma, Norman, Oklahoma
CORRESPONDING AUTHOR: V Lakshmanan, 120 David L. Boren Blvd., Norman, OK 73072
E-mail: lakshman@ou.edu

DOI:10.1175/2010BAMS2870.1

egory. Instead, the contestants submitted the result of their AI model to the competition chairs, who scored each submitted result against the true classification. Because the AI models were scored on an independent dataset, this was a fair test of generalization in a real-world meteorological scenario.

The conventional wisdom in the meteorology–AI community going into the competition that year was that the choice of AI techniques mattered a huge deal. In fact, the members of the STAC had collaborated on writing a book, titled *Artificial Intelligence Methods in the Environmental Sciences*, that laid out different ideas on when to choose among the various AI models.

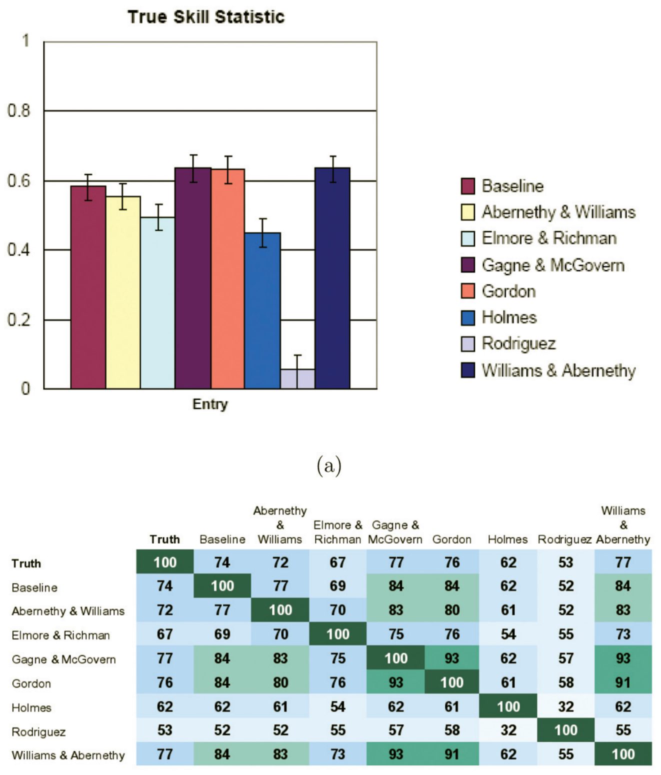
When applied in a blind comparison on a real-world meteorological dataset, however, we discovered that the choice of AI technique did not matter much. Once features had been computed from the dataset, pretty much any modern AI technique—neural networks, decision trees, random forests—all performed quite similarly. Statistically, the performances of the top three entries were indistinguishable—when presented with a set of inputs, the techniques would nearly always provide the same answer. It was impossible, based on just the output of the techniques, to identify which was which (See Fig. 1).

In other words, based on these data, it could be concluded that, as a research community, our habit of picking a problem and selecting an approach was not ideal. It didn't matter much whether the final AI model was a neural network or a decision tree: the performance would be quite similar. We would need to devote the maximum amount of care to the formulation of the problem—to the creation of features in the dataset.

Indeed, the winning entry that year combined several of the features in the provided dataset based on a knowledge of the shortcomings of radar tracking algorithms (Williams and Abernethy 2008). Because this combined statistic was better behaved (in the sense that it was less likely to be subject to radar sampling errors) than the underlying individual statistics, the AI model (a random forest, as it turned out) that used the combined feature outperformed the rest. A fortuitous coincidence led to this conclusion—the best student entry (Gagne and McGovern 2008) also used a random forest, but without the combined variable. The difference in performance between the winning entry (Williams-Abernethy) and the student entry (Gagne-McGovern) could be attributed wholly to the incorporation of the new variable.

The first year of the competition, we had begun to form a consensus that most of our effort in applying AI to the environmental sciences would have to be in formulating the features that fed into whichever AI model was selected. The actual AI model chosen was secondary in achieving skill.

Of course, the consensus that the particular AI method did not matter was subject to common-sense caveats such as understanding the data and not overfitting. The entries with poor performance in the first year's competition either got the relative frequencies of the categories wrong or chose an imputation method that ignored what was known about the dataset.



(a)

	Truth	Baseline	Abernethy & Williams	Elmore & Richman	Gagne & McGovern	Gordon	Holmes	Rodriguez	Williams & Abernethy
Truth	100	74	72	67	77	76	62	53	77
Baseline	74	100	77	69	84	84	62	52	84
Abernethy & Williams	72	77	100	70	83	80	61	52	83
Elmore & Richman	67	69	70	100	75	76	54	55	73
Gagne & McGovern	77	84	83	75	100	93	62	57	93
Gordon	76	84	80	76	93	100	61	58	91
Holmes	62	62	61	54	62	61	100	32	62
Rodriguez	53	52	52	55	57	58	32	100	55
Williams & Abernethy	77	84	83	73	93	91	62	55	100

(b)

Fig. 1. The particular AI technique used did not matter much: winning entries in the 2007–08 competition were quite closely clustered together. (a) Skill scores of the different entries, with a 95% confidence interval marked. (b) The better-performing entries were statistically indistinguishable, yielding the same answers (whether correct or incorrect).

THE 2008–09 COMPETITION DATA. For the second year (2008–09), we chose to use a dataset gathered by the National Severe Storms Laboratory Winter Precipitation Identification Near the Ground (W-PING) experiment, which is still ongoing. The classification task was to use polarimetric radar data, collected with the KOUN radar, along with limited environmental information, to develop a hydrometeor classification algorithm that would distinguish between frozen and liquid hydrometeors, or none. In W-PING, the public is asked to observe winter precipitation in situ and enter their observation on a Web site, distinguishing between the following categories: rain, drizzle, freezing rain, freezing drizzle, ice pellets (sleet), graupel, snow, hail, and none, all within a 150-km radius from the KOUN radar. Since a cold-season hydrometeor classification algorithm must be able to distinguish between frozen, liquid, and no precipitation, the above categories were amalgamated into the three used in the competition. Freezing rain and freezing drizzle were combined with rain and drizzle, and classed as "liquid." Snow, ice pellets (sleet), graupel, and hail were all combined into "frozen," while "none" was retained as is.

The observed precipitation-type data were quality controlled using rather broad criteria. If an observation was clearly inconsistent with nearby observations in time and space, it was removed. For example, observations of "hail" in the midst of "snow" were removed. Observations well outside of the project area were removed, as were obvious duplicates. Around each ground observation, radar data for each polarimetric radar parameter (such as specific differential phase, Kdp) were averaged over a 5 x 5 (range by azimuth) kernel centered on each ground observation. Only observations associated with radar data between 0.3 km and 1.2 km AGL were used. Within that height range, only the lowest scan was chosen. All data were filtered to remove observations within ground clutter. The data were, however, prefiltered in the sense that all of the included variables are a priori reasonable choices as predictors. No "distractor" variables were included that were at best useless and at worst expected to negatively affect the performance of a technique that was not carefully crafted by someone understanding the meteorological problem to some extent.

Data were taken from three main events for which about 2,650 observations were initially logged. After the rudimentary quality control, about 2,500 remained. Of these 2,500, 1,573 met all the other criteria stated earlier. It is important to note that, until the time of the competition, these data remained unique and had not been distributed anywhere, and so were unavailable from any other source. Hence, no one outside of the W-PING project had any access whatsoever to these data.

The testing data was generated by sampling, without replacement, from the full dataset. The testing data constituted 30% of the full dataset, leaving the other 70% for training. No attempt was made to "balance" the proportion of the various categories. The training data contained 58.3% frozen, 28.2% liquid, and 13.5% none, while the testing data contained 56.7% frozen, 32.9% liquid, and 11.3% none.

TRUST, BUT VERIFY

The Web page used for the reference to the Peirce Skill Score (PSS) contained a typographic error that had been overlooked for years (it has since been corrected). The error became apparent when one entrant submitted a classifier that was admittedly "hedged," submitted with the belief that this formulation would result in a very high score. We were surprised by this, because the PSS was chosen specifically because it is not subject to hedging. That particular entry scored poorly. The entrant was bewildered at the skill score his entry achieved and, based on his analysis of the (erroneous) PSS formula, found that this value was not one of the numerically possible scores his entry could have received. After conferral between the competition chairs and the contestant, it was discovered that the referenced Web page possessed the error. The correct formulation for the multicategory PSS is shown in the figure. The variables *i* and *j* here are equal and refer to the number of categories. The erroneous score had y_i in the denominator instead of o_i.

The erroneous formulation leads to a score that is easily hedged and typically provides slightly higher values than does the correct PSS formulation. The session chairs conferred with the AI STAC and decided to use the correct score for rankings, and to use the unfortunate situation as a reminder to always trust, but verify.

$$PSS = \frac{\sum_{i=1}^{I} p(y_i, o_i) - \sum_{i=1}^{I} p(y_i)p(o_i)}{1 - \sum_{j=1}^{I} [p(y_i)]^2}$$

Wrong formula

$$PSS = \frac{\sum_{i=1}^{I} p(y_i, o_i) - \sum_{i=1}^{I} p(y_i)p(o_i)}{1 - \sum_{j=1}^{I} [p(o_i)]^2}$$

Correct formulation

Fig. SB1. The Web page that contestants were pointed to had a typo in the formula for the Peirce Skill Score.

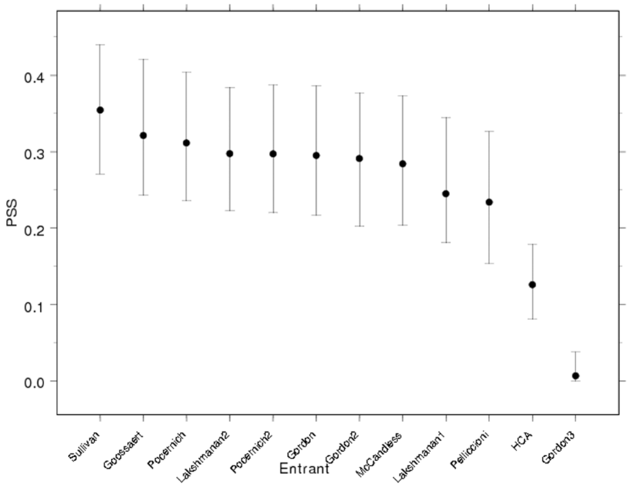


Fig. 2. In the 2008–09 competition, there was no significant difference in the Peirce Skill Score (PSS) of the better performing entries. This was further verified by a bootstrapping test, as explained in the text.

In the second year, we chose Peirce's Skill Score (PSS) for determining the winners. The PSS is a multicategory skill score and is therefore amenable to a three-category classification problem. It is also equitable and so not subject to hedging or gaming (creating forecasts that do not represent the true beliefs of the developer). Contestants were provided a Web page[1], a product of the Joint Working Group on Forecast Verification Research, for the formulation of the PSS.

Figure 2 shows the resulting entrants' scores along with 95% confidence intervals for those scores based upon bootstrap resampling and bootstrap tilting. As was the case for the 2007–08 competition, there was no significant difference between the various entries. The entry by Sullivan had the highest PSS (although by a statistically insignificant amount) and was deemed the winner. The lack of statistical significance is not so much due to shortcomings of particular methods as it is to natural variability in the dataset. Without any additional constraints, based on the apparent natural variability, it seems prudent to use whatever method is most easily understood by end users.

Hydrometeor classification algorithms have been the subject of extensive research, with previous classification methods ranging from the use of pattern recognition to fuzzy logic to rules based on conceptual models. The machine intelligence approach pioneered in the competition is a fundamentally different way to address hydrometeor classification. In particular, it relies, first and foremost, on building up a dataset of observations. In other words, it is a data-driven approach and is particularly apt because the observations are of hydrometeors aloft, but the forecasting problem is to predict hydrometeors on the surface. As observing instruments become better and observing networks become denser, a data-driven approach will become feasible for many more problems in the atmospheric sciences.

SUMMARY. The AMS Committee on Artificial Intelligence Applications to the Environmental Sciences has held seven sessions since its inception. Committee members are a mix of atmospheric scientists, engineers, and computer scientists. Despite the breadth of expertise, members all share a keen interest in AI techniques and seek to engage the broader atmospheric science community to illustrate how AI techniques can help solve real-world problems. The committee decided that an AI contest was an ideal venue to connect theorists, practitioners, and industry in a meaningful dialogue. The spirited discussions at the two contests suggest that the format has been a success in engaging the audience. At times, the findings have been counterintuitive to conventional wisdom; however, they have exposed our community to a wide array of methodologies causing us to reconsider the philosophy of attacking a problem. Such interactions have motivated many participants to augment their tool kits of favorite AI methods. In doing so, we all grow and benefit from the richness of a wider perspective.

ACKNOWLEDGMENTS. Funding for the authors was provided under NOAA-OU Cooperative Agreement NA17RJ1227.

FOR FURTHER READING

Elmore, K., and M. Richman, 2009: The 2008 artificial intelligence competition data: Source and characteristics. *7th Conf. on Artificial Intelligence Applications to Environmental Science*, Phoenix, AZ, Amer. Meteor. Soc., 2.1.

[1] www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html whose content has now migrated to www.cawcr.gov.au/projects/verification/.

Gagne, D. J., and A. McGovern, 2008: Using multiple machine learning techniques to improve the classification of a storm set. *6th Conf. on Artificial Intelligence Applications to Environmental Science*, New Orleans, LA, Amer. Meteor. Soc., 2.4.

Guillot, E., T. Smith, V. Lakshmanan, K. Elmore, D. Burgess, and G. Stumpf, 2008: Tornado and severe thunderstorm warning forecast skill and its relationship to storm type. *24th Conf. on IIPS*, New Orleans, LA, Amer. Meteor. Soc., 4A.3.

Haupt, S., A. Pasini, and C. Marzban, Eds., 2008: *Artificial Intelligence Methods in the Environmental Sciences*. Springer, 424 pp.

Hong, Y., K. Hsu, S. Sorooshian, and X. Gao, 2004: Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *J. Appl. Meteor.*, **43**, 1834–1853.

Kessinger, C., S. Ellis, and J. Van Andel, 2003: The radar echo classifier: A fuzzy logic algorithm for the WSR-88D. *3rd Conf. on Artificial Applications to the Environmental Sciences*, Long Beach, CA, Amer. Meteor. Soc.

Krasnopolsky, V. M., M. Fox-Rabinovitz, H. Tolman, and A. A. Belochitski, 2008: Neural network approach for robust and fast calculation of physical processes in numerical environmental models: Compound parameterization with a quality control of larger errors. *Neural Networks*, **21**, 535–543.

Lakshmanan, V., 2009: The simpler the better. *6th Conf. on Artificial Applications to the Environmental Sciences*, Phoenix, AZ, Amer. Meteor. Soc., 3.5.

——, and T. Smith, 2009: Data mining storm attributes from spatial grids. *J. Atmos. Oceanic Technol.*, **26**, 2353–2365.

——, E. Ebert, and S. Haupt, 2008: The 2008 artificial intelligence competition. *6th Conf. on Artificial Intelligence Applications to Environmental Science*, New Orleans, LA, Amer. Meteor. Soc., 2.1.

Liu, H., and V. Chandrasekar, 2000: Classification of hydrometeors based on polarimetric radar measurements: Development of fuzzy logic and neuro-fuzzy systems, and in situ verification. *J. Atmos. Oceanic Technol.*, **17**, 140–164.

Marzban, C., and G. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.*, **35**, 617–626.

——, and V. Lakshmanan, 1999: On the uniqueness of Gandin and Murphy's equitable performance measures. *Mon. Wea. Rev.*, **127**, 1134–1136.

Mecikalski, J., K. Bedka, S. Paech, and L. Litten, 2008: A statistical evaluation of GOES cloud-top properties for nowcasting convective initiation. *Mon. Wea. Rev.*, **136**, 4899–4914.

Park, H., A. Ryzhkov, D. Zrnic, and K. Kim, 2009: The Hydrometeor Classification Algorithm for the polarimetric WSR-88D: Description and application to an MCS. *Wea Forecasting*, **24**, 730–748.

Peirce, C., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Rahman, M., R. Jacquot, E. Quincy, and R. Stewart, 1981: Two-dimensional hydrometeor image classification by statistical pattern recognition algorithms. *J. Appl. Meteor.*, **20**, 536–546.

Scharfenberg, K. A., and Coauthors, 2005: The Joint Polarization Experiment: Polarimetric radar in forecasting and warning decision making. *Bull. Amer. Meteor. Soc.*, **20**, 775–788.

Trafalis, T., A. White, B. Santosa, and M. Richman, 2002: Data mining techniques for improved WSR-88D rainfall estimation. *Comput. Ind. Eng.*, **43**, 775–786.

Williams, J., and J. Abernethy, 2008: Using random forests and fuzzy logic for automated storm type identification. *6th Conf. on Artificial Intelligence Applications to Environmental Science*, New Orleans, LA, Amer. Meteor. Soc., 2.2.