

Reaching Scientific Consensus Through A Competition

Valliappa Lakshmanan^{1,2*}, Kimberly L. Elmore^{1,2}, Michael B. Richman³

*Corresponding author: V Lakshmanan, 120 David L. Boren Blvd, Norman OK 73072; lakshman@ou.edu

¹Cooperative Institute of Mesoscale Meteorological Studies, University of Oklahoma; ²National Oceanic and Atmospheric Administration / National Severe Storms Laboratory; ³School of Meteorology, University of Oklahoma

1. Why a Competition?

In a manner similar to the other Science and Technology Advisory Committees (STACs), the AMS' 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 (Trafalis et al. 2002; Hong et al. 2004), nowcasting lightning (Lakshmanan and Smith 2008), predicting convective initiation (Mecikalski et al. 2008), diagnosing tornado probability (Marzban and Stumpf 1996), controlling radar data quality (Kessinger et al. 2003), and approximating computationally expensive models (Krasnopolsky et al. 2008).

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 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 a dataset and explain the dataset. Then, a variety of researchers would apply different techniques to the AI data set. At the conference, the data set 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 ranked on a predetermined measure of skill and certificated distributed, 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.

2. Anything Goes

The first year (2007/2008), 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 extracted by the storm tracking method of Lakshmanan and Smith (2008). These attributes, as well as the human categorization of the storms by type, were provided

to the contestants. The contestants used this training data set to create their AI models. A separate test data set consisting of a similar variety of storms was later provided to the contestants but for the test dataset, the human storm type category of the storms was withheld. 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 is 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 (Haupt et al. 2008) 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 data set, pretty much any modern AI technique – neural networks, decision trees, random forests – all performed quite similarly (Lakshmanan et al. 2008). Statistically, the performances of the top three entries were indistinguishable – when presented 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 Figure 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 Abernathy 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 also used a random forest but without the combined variable – the difference in performance between the winning entry and the Gagne-McGovern entry 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 (Lakshmanan et al. 2008).

3. Trust But Verify

For the second year (2008/2009), we chose to use a data set 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 (Scharfenberg et al. 2005), 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 website, 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 HCA 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 observation 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 obviously duplicate entries. Around each ground observation, radar data for each polarimetric radar parameter (such as 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.

Data were taken from three main events for which about 2650 observations were initially logged. After the rudimentary quality control, about 2500 remained. Of these 2500, 1573

met all the other criteria stated earlier. It is important to note that these data are unique in that no additional such data exist from any 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 data set. The testing data constituted 30% of the full data set, leaving the other 70% for training. No attempt was made to "balance" the proportion of the various categories. The training data contain 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.

The second year, we chose Peirce's Skill Score (PSS) for determining the winners (Elmore and Richman 2009). The PSS is a multi-category skill score and is therefore amenable to a 3-category classification problem. It is also equitable (Marzban and Lakshmanan 1999) 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. However, this web page contained a typographic error that had been overlooked for years (it has since been corrected: see Figure 2). The error became apparent was when one entrant (Gordon) 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 (Gordon 3: see Figure 3) scored poorly. Gordon 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 "possible" scores his entry could have received. After conferral between the competition

¹http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

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 Figure 2b where $I = J =$ number of categories. The erroneous score had f_i in the denominator instead of o_i (See Figure 2a).

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.

Figure 3 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 2007-2008 competition, there was no significant difference between the various entries, save for Gordon 3. The lack of statistical significance is not so much due to shortcomings of particular methods as it is to natural variability in the data set. Without any additional constraints, based on the apparent natural variability, it seems prudent to use whatever method is most easily understood by end users (Lakshmanan 2009).

4. 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 the 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 toolkits of favorite AI methods. In doing so, we all grow and benefit from the richness of a wider perspective.

Acknowledgements

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

REFERENCES

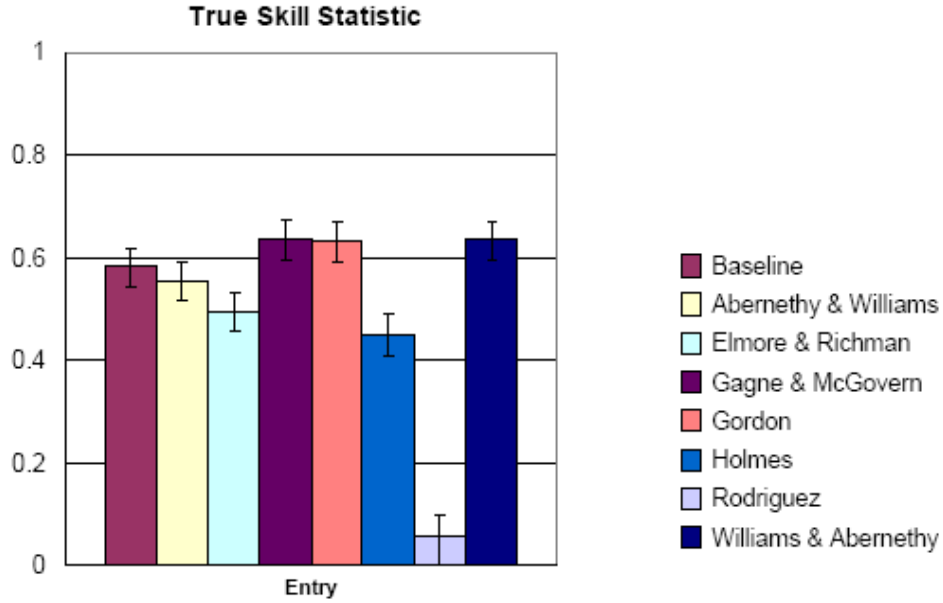
- Elmore, K. and M. Richman, 2009: The 2008 artificial intelligence competition data: Source and characteristics. *7th Conference on Artificial Intelligence Applications to Environmental Science*, Phoenix, Amer. Meteor. Soc., 2.1.
- Guillot, E., T. Smith, V. Lakshmanan, K. Elmore, D. Burgess, and G. Stumpf, 2008: Tornado and severe thunderstorm warning forecast skill and its relationship to storm type. *24th Conference on IIPS*, New Orleans, 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 Conference 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 Conference on Artificial Applications to the Environmental Sciences*, Phoenix, AZ, Amer. Meteor. Soc., 3.5.
- Lakshmanan, V., E. Ebert, and S. Haupt, 2008: The 2008 artificial intelligence competition. *6th Conference on Artificial Intelligence Applications to Environmental Science*, New Orleans, Amer. Meteor. Soc., 2.1.
- Lakshmanan, V. and T. Smith, 2008: Data mining storm attributes from spatial grids. *J. Ocea. and Atmos. Tech.*, Under review.
- Marzban, C. and V. Lakshmanan, 1999: On the uniqueness of gandin and murphy’s equitable performance measures. *Monthly Weather Review*, **127 (6)**, 1134–1136.

- Marzban, C. and G. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. App. Meteor.*, **35** (5), 617–626.
- Mecikalski, J., K. Bedka, S. Paech, and L. Litten, 2008: A statistical evaluation of GOES cloud-top properties for nowcasting convective initiation. *Monthly Weather Review*, **136** (12), 4899–4914.
- Scharfenberg, K. A., et al., 2005: The Joint Polarization Experiment: Polarimetric radar in forecasting and warning decision making. *Bulletin of the American Meteorological Society*, **20**, 775–788.
- Trafalis, T., A. White, B. Santosa, and M. Richman, 2002: Data mining techniques for improved WSR-88D rainfall estimation. *Computers and Industrial Engineering*, **43**, 775–786.
- Williams, J. and J. Abernathy, 2008: Using random forests and fuzzy logic for automated storm type identification. *6th Conference on Artificial Intelligence Applications to Environmental Science*, New Orleans, Amer. Meteor. Soc., 2.2.

List of Figures

1	The particular AI technique used did not matter much: winning entries in the 2008 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).	12
2	The web page that contestants were pointed to had a typo in the formula for the Peirce Skill Score.	13
3	In the 2008-2009 competition, there was no significant difference in the Pierce Skill Score (PSS) of the better performing entries. This was further verified by a bootstrapping test, as explained in the text.	14



(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 2008 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).

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

Wrong formula

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

Correct formulation

FIG. 2. The web page that contestants were pointed to had a typo in the formula for the Peirce Skill Score.

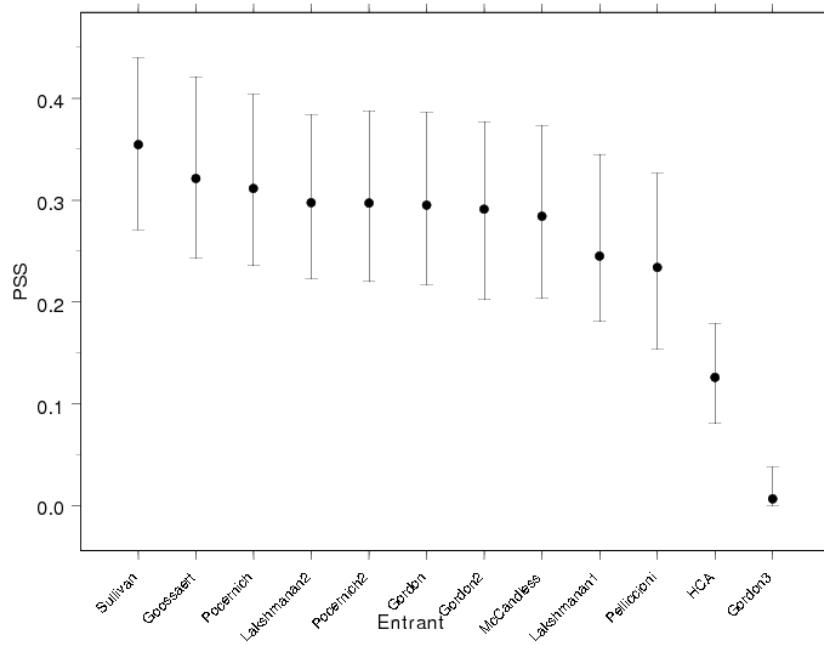


FIG. 3. In the 2008-2009 competition, there was no significant difference in the Pierce Skill Score (PSS) of the better performing entries. This was further verified by a bootstrapping test, as explained in the text.