NOTICE

**CERTAIN DATA
CONTAINED IN THIS
DOCUMENT MAY BE
DIFFICULT TO READ
IN MICROFICHE
PRODUCTS.**

Conf-9010289-4

ANL-HEP-CP--90-118

DE91 006489

Believability of Signals from Cosmic Ray Sources*

by Maury Goodman,
Argonne National Lab. Argonne Ill. 60439

JAN 11 1991

Abstract

This paper discusses some of the criteria by which an observer judges whether to believe a signal or limit that has been reported for a cosmic ray source. The importance of specifying the test before looking at the data is emphasized.

Introduction

At conferences and seminars, we have often heard speakers make statements like the following: "This result may be evidence for cosmic ray emission from object X. The chance that we would get this result as a statistical fluctuation of the background is 10^{-3} ." The following are possible alternative implications of the statement:

1. Object X is a source of the cosmic rays.
2. The result was a very unlikely statistical fluctuation.
3. The result was due to a systematic effect that wasn't properly taken into account.
4. The chance probability of a statistical fluctuation was not properly calculated.

A large number of observations of many sources have been presented at various conferences and throughout the literature.¹ A variety of techniques have been used, but a consistent interpretation of the various results has failed to emerge. One possibility is that the sources are real, although they may be sporadic and involve new astrophysical and particle phenomena.

Certain features and potential inconsistencies of some of the results have led many to be skeptical about most or all of the reported sources. Many signals reported at conferences are never published. Null observations are often not presented publicly. This has led to a certain cynicism in the field, leading to the disregarding of potentially significant results as "not significant enough".

This paper's goal is to discuss the framework in which we evaluate such results and then believe or disbelieve them. I feel that many results are due to the use of improper statistical procedures. Section I is a list of some of the characteristics that many of us expect from a believable signal. Section II looks into the dangers of a-posteriori analysis techniques. Section III covers some of the characteristics of a doubtful result. In Section IV some of the issues in setting limits based on null observations are considered.

MASTER
DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

*Submitted to the Proceedings of the International Conference on High Energy Gamma Ray Astronomy, Ann Arbor, MI

Work supported by the U.S. Department of Energy, Division of High Energy Physics, Contract W-31-109-ENG-38

I. Characteristics of a real signal

High Energy cosmic ray sources with a wide variety of possible signatures have been reported. Besides dc excess and phase excess, there have been reports of bursts of cosmic ray emission over a large number of time scales from many potential sources. While reading a paper or listening to a presentation, one often asks oneself, "Do I believe this result?" I discuss here six criteria that are commonly used in answering this question: 1.) significance, 2.) repeatability, 3.) control of systematics, 4.) self-consistency, 5.) consistency with others, and 6.) a-priori hypothesis tested. These characteristics are neither necessary nor sufficient to have a believable result. The ν observations of SN1987A failed some of these concerns.²⁾ Yet the observation is generally believed. The Whipple observation of the Crab is generally accepted, even though no other group has been able to confirm the azimuth analysis result³⁾. Other observations of various sources, however, fail more of these criteria. I believe 6.) may be the most important and least satisfied of the group. It is discussed in detail in the next section; here I review the other points:

1. Is the result statistically significant? We often speak of the probability P that this result may have happened by chance. It is likewise often asked "How many σ ?" (most physicists should be aware that their intuitive understanding of σ is only for Gaussian distributions. By a 3σ effect, we usually mean a chance probability corresponding to a 3σ effect for a Gaussian distribution. It is sometimes forgotten that 3 times the 1σ effect for a non-Gaussian distribution can have a very different chance probability.) The choice of the level of significance is arbitrary and depends on the circumstances. Using a level of significance of 5%, one would be wrong one time in 20. A " 3σ " effect corresponds to a chance probability of 1.4×10^{-3} (using one side of the Gaussian distribution).
2. Repeatability is normally regarded as an important aspect of physics measurements. Many of the reported signals have in fact not been repeated in subsequent observations by the same group. This is often cited as evidence for sporadic emission. It would be desirable to see sporadic emission of more than one source, or more than once from the same source.
3. Each experiment needs to understand its own systematic effects. Once they have been identified, they can usually be studied or estimated with the data.
4. Each experiment should do all possible consistency checks on its own data. For example, if an experiment measures a phase excess or power, is the dc excess compatible with the required number of events? Also are data from background directions consistent with assumptions about time uniformity?
5. Finally, different experiments measuring the same phenomena should be consistent. If a sporadic event has been recorded, this requires simultaneous observation. If two experiments observe the same source, it greatly lowers the potential for systematic error. However it does not eliminate the possibility, and both groups need to use sound statistical methods. Care must also be taken in the statistical combination of two results⁵.

II. A-Priori vs. A-posteriori analysis

I advocate that levels of statistical significance be quoted only for tests that were formulated before the data is examined. Results which are due to tests formulated after "looking at the data" should be presented only as hypotheses to be tested with new data sets or by other groups.

It has sometimes been assumed that one can in an unbiased way decide what question to ask after the data have been analyzed. The history of physics includes many important discoveries made in this way. An observation conducted this way of a transient effect (and therefore difficult to confirm) is fraught with great danger, however. While examining data, the careful analyzer is drawn to unusual aspects of the data. In noting any such particular aspect, it is usually impossible to estimate how many other unusual effects could have appeared just as significant. These can be considered hidden trials, which need to be taken into account, yet the number of such trials simply isn't known. Once an unusual effect is seen, a large number of plots are usually examined, most of which are highly correlated. Again, unless the analysis is specified in advance, it is usually impossible to assess the true statistical significance of the effect.

A-posteriori effects, no matter how significant they appear, should only be regarded as new hypotheses requiring further test. Chance probabilities are not likely to have a quantitative meaning for such effects. Therefore they should not be quoted at all, or should only be quoted with clear caveats that they are different from chance probabilities for a-priori hypotheses.

It should be emphasized, however, that a-posteriori data analysis should be done. One reason for doing an experiment is that we don't know what new effects to expect. Intuition plays a great role in data analysis, and seemingly unusual effects in the data need to be studied thoroughly. It is the attribution of a quantitative chance probability for such an effect to which I object.

I am certainly not the first to point out the importance of testing a-priori and not a-posteriori hypotheses. Experimenters have continually violated this rule because so many hypotheses are formulated a-posteriori and one wants to get as much out of the data as possible. I reiterate the problem by quoting Wallis⁴⁾; "An investigator who after inspecting the data decides what to test or how to make the test can, by virtue of the fact that any sample has unique characteristics, disprove any hypothesis."

It is sometimes difficult to *remember* what a-priori hypotheses were. In the social sciences, we are aware of the necessity of double-blind experiments in certain situations. One thing to beware of is the changing hypothesis. Variability is such an example. Early data sets were searched for dc emissions of sources and many signals were found. Let's say the null hypothesis was rejected with a significant chance probability P_s . Other experiments have generally contradicted these observations (except for the Crab in VHE γ -rays). The possibility certainly exists that the source was variable in time, but this a different hypothesis than the one originally tested. Therefore it is not appropriate to use P_s as

the probability for this new hypothesis. Again in this case, it is inappropriate to use any chance probability, and one is left with an interesting phenomena suggested by the data.

Another problem may be hypotheses that are suggested after a partial examination of the data. The analyzer may suggest (and believe) that the analysis that gave the unusual result was the a-priori hypothesis, when in fact it wasn't. This is almost impossible for the outside observer to judge, but as a rule of thumb, I suggest that, *"If the question isn't obvious, beware of the answer"*. An "a-priori" hypothesis that seems contrived may reflect a large number of hidden trials for which the chance probability should be modified.

III. Characteristics of a doubtful signal

The obvious characteristics of a doubtful signal are the antithesis of the characteristics of a believable signal. It may be statistically insignificant, based on an a-posteriori test, not repeatable, influenced by uncontrolled systematics, self inconsistent and/or contradicted by others.

Another characteristic of many such effects is their observation soon after a detector starts to accumulate data. Analysis techniques are being developed, and creative physicists invariably turn their attention to unusual aspects of the data. If such an effect fails to repeat in subsequent data that is an indication that the effect only appeared unusual because hidden trials were not taken into consideration.

The physics implication of a result certainly has some bearing on its believability. A 3σ result on the moon shadow is more likely to be accepted than a 3σ result which implies 2 new particles. This is not a statistical statement, of course, but a general judgement of the 4 possibilities mentioned at the beginning of this paper. In principle, we should set our required level of significance in advance. In practice, this is less important because we are rarely forced to decide if a given result is believed or not. Some groups may have a policy that a signal from a new source may require $n\sigma$ before they would write a paper. If there is a well-defined a-priori hypothesis that is being tested, any result of any level of significance should be presented publicly and it stands on its own. If a result is based on a-posteriori hypothesis, this paper suggests the $n\sigma$ criterion is without quantitative meaning, so "going public" with the effect as a new hypothesis is based on other issues.

IV. Consistency between methods used to find signals and set limits

A truly calculated 3σ effect will happen by chance only 1.4 times in 1000. Yet most of us are aware of claimed 3σ effects which "went away". As a consequence, some groups say that only a 5σ or more result should be considered as evidence for a source. I call this the "discount factor" and I feel it may be inappropriate.

In the absence of a signal, it is desirable to quote upper limits on the flux, which may then be compared to the flux calculated for a signal from another

experiment. A 90%CL upper limit implies, in a sense, the absence of a 1.3σ effect. Mathematically, an upper limit can be calculated for any data, but I feel that quoting only a 90%CL upper limit when there is a 2.7σ effect, for example, is potentially misleading. A 5σ upper limit would be at 99.99997%CL and would be higher.

Should one have higher standards for the reporting of a signal than for a limit? I think the answer is no. One may want a high confidence level before accepting the existence or sources (or any other new physics result). But then one should want an equally high confidence level before disproving them. One may disbelieve some experiment based on conventional wisdom. That should not lead to acceptance of a different standard for experimental evidence needed to contradict the results.

Summary

This paper is one physicist's view of a rather controversial subject-the possible existence of cosmic ray signals from discrete astrophysical objects. The large number of significant yet confusing reports of source observation may be the result of improper calculation of chance probabilities. The largest contribution to this may be the use of tests based on a-posteriori effects.

A common reaction to the confusing situation, however, has been the discounting of potentially significant results as "insignificant". This is not a proper response. We should restore the dignity of the 3σ effect.

Acknowledgements

I want to thank Steve Biller, J. Uretsky, Tom Fields, and Trevor Weekes for helpful comments. All presumptuous statements are the author's own.

References

1. D. Fegan, to be published in proceedings of the 21st ICRC, **11**, Adelaide, (1990). Also see D. Nagle, T.K. Gaisser and R. J. Protheroe, Annual Reviews of Nuclear and Particle Science **38**, (1988) and J.-M. Bonnet-Bidaud and G. Chardin, Physics Reports **170**, 325 (1988).
2. R. M. Bionta et al., Phys. Rev. Lett., **58** 1494 (1987). and K. Hirata et al., Phys. Rev. Lett., **58** 1490 (1987).
3. T.C. Weekes et al., Ap. J., **175**, 379 (1989).
4. W. A. Wallis, Econometrica **10**, 229 (1942).
5. M.C. Goodman, "Signals from Cosmic Ray Sources. Some Statistical Issues" submitted to the Proceedings of the 1990 Snowmass Summer Study; Jack Uretsky, "Combining Independent Tests of Significance and Why Would anyone Want to?", PDK-441, Argonne, 1990 (unpublished); and D.A. Lewis, in Arkansas Conference Proceedings: Nuclear Physics B, p299, 1989, and D.A. Lewis, Astronomy and Astrophysics **13** 696(1990).

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

-END-

DATE FILMED

01 / 29 / 91