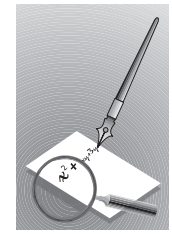
commentary and analysis



On the Lack of Accountability in Meteorological Research

1. Introduction

There are three primary means by which scientific research is most commonly evaluated. One is by review and competition of proposals, another is by questioning in public presentations, and the third is by review of publications, both before and after their dissemination. The last includes the scientific process by which new experiments are performed to support or reject proposed theories. Indeed, the peer review process is considered sacrosanct in science.

All researchers and their managers will admit that the evaluation processes do not work perfectly. Of course, mistakes will be made, usually unintentionally and rarely otherwise. It is my contention, however, that our evaluation processes are currently functioning so poorly that the integrity of the science and its timely progress are actually being jeopardized. I will attempt to justify this contention by describing my observations of the functioning of the above three means of scientific evaluation, starting with the last. I will refrain, however, from providing details regarding the observations provided for the sake of brevity. Examples will be provided if requested (NCAR, P. O. Box 3000, Boulder, CO 80307; rmerrico@ucar.edu).

2. Publication reviews

Too frequently, published papers contain fundamental errors. Some are the kinds of errors that even scientists who are not experts in the particular topic should catch. Others require an expert in a particular area, and although the error may be fundamental, it will remain undisclosed unless one such expert reveals it. Still others may be characterized as simply errors in basic scientific procedure, that is, failure to look at a result or to consider contradictions with earlier studies brought to the investigator's attention.

The presentation in many papers is careless. Units or constants may not be provided. Grid resolution may not be given. Comparisons may be made between two schemes without one of the schemes being adequately described. Definitions may be absent. Preprocessing, postprocessing, or scaling factors may be applied but not described. It is not just ancillary information that is lacking, but rather critical details necessary for properly interpreting the primary results. Without such information, in fact, composing scientific reviews is often rendered extremely difficult or even impossible: How can a piece of work be adequately evaluated or duplicated if what was really done or meant is not adequately stated?

Some papers abound in unsupported claims stated as facts. Supporting references, if cited, are sometimes used out of proper context. Statements and results may even be absurd. Unfortunately, once in print, unsupported claims tend to become cited as factual.

The unnamed papers to which I am referring are not obscure articles. Instead they are often cited and used to motivate other projects. Both editors and authors have told me that some of these articles have sailed through the review process. Some have even been garnished with awards, although their fundamental errors render them worse than useless: They would simply be useless if they were not subsequently cited in defense of continued bad science.

One objective analysis is pertinent here. Figure 1 presents the annual numbers of articles and comments published during 1976–98 for two American Meteorological Society (AMS) journals. There is much year-to-year fluctuation, but the number of comments has diminished by approximately one-half during this period. The number of articles has generally risen, however, so that the percentage of papers commented upon has declined by more than one-half. Is this decline because the paper quality is generally improving? Is it because comments are instead being incorporated in

longer articles? Since it does appear to me that quality is actually degrading and since I have not seen significant critical review in articles, I expect that neither of these explanations is correct and instead that the dearth of comments is due to other factors.

One difference between the meteorological and some other scientific communities is its small size. This may partially explain the dearth of comments in the literature. The same people are often reviewing each other's works, and criticism can easily be returned. Many researchers are actually friends who are likely reluctant to criticize in print. I personally enjoy the friendly atmosphere, but it is unhealthy for the science if debate is thereby discouraged.

I have used the terms "sometimes," "may," "some," etc., in my comments here. If these vague words referred to less than 20% of papers, perhaps that would be acceptable. Some poor papers would get through any review process. In research topics that I read, however, 50% may be a closer estimate of misleading or fundamentally wrong papers. Such a number is not easy to evaluate. If it is this high, however, we have a problem that should not be ignored.

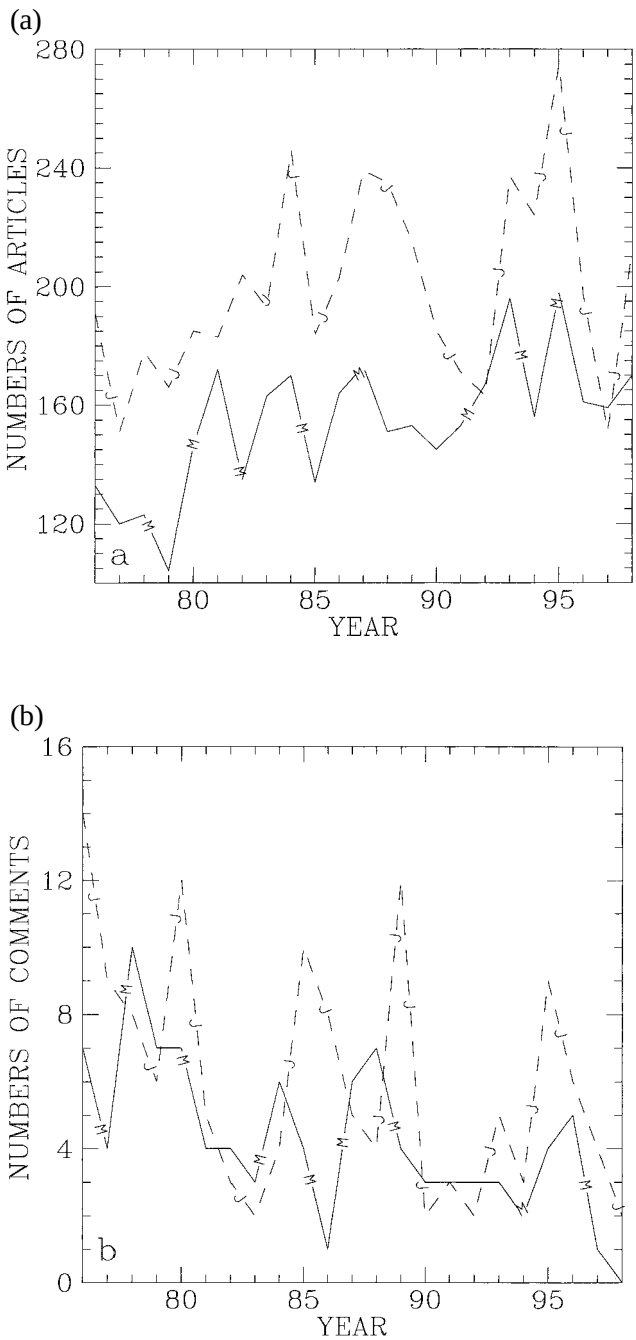


FIG. 1. Numbers of (a) articles and (b) comments in *Monthly Weather Review* (solid line) and the *Journal of the Atmospheric Sciences* (dashed line) within each indicated year.

3. Meetings

Everyone has attended meetings at which questioning or commenting directly after presentations was disallowed, usually because speakers have gone beyond their allotted time. At some meetings, even those called workshops, however, questioning and commenting is discouraged at the onset by design. I think the price paid for the absence of more effective audience participation has been grossly underestimated!

Why is questioning and commenting so critical? First, usually few people in the audience are sufficiently knowledgeable to evaluate a presentation. When questioning has been permitted, I sometimes have learned that what sounded scientifically accurate was not, or vice versa. In the latter cases, someone in the audience has corrected misleading impressions left by the presentation. Second, comments can put the work in a wider context and offer additional perspectives, especially if the speaker is a novice and more expert scientists are in the audience.

I have witnessed talks in which fundamental errors were made, with no real questioning occurring, and then a year later heard similar talks by the same speakers having the same errors! The second talks were not pulled out of dusty files; they described continuing research. The speakers had wasted considerable intervening time by being deprived of an opportunity to correct the errors after the first talks.

Speakers could be questioned or critiqued privately rather than in public. Sometimes this never occurs because it is assumed another colleague will do so or there is never a good opportunity. Likely,

some potential questioner thinks, "This seems wrong, or I do not understand this, but I do not want to appear ignorant," and therefore refrains from commenting even if an opportunity exists. Often others are thinking similarly, however, and they too could have learned from the comment and its reply. An advantage of questioning and commenting in a public forum is that others are present who can be drawn into a dialogue. Maybe someone else knows the proper follow-up question or can answer a question the speaker cannot. When else will so many experts in the field likely be together and have the context of the presentation focusing their attention?

Many meetings are so large, with so many submitted papers, that it is difficult to permit extensive questioning without severely limiting the number or duration of presentations. At large conferences, people also tend to be dispersed in different conference rooms or maybe even different hotels, so that finding someone to discuss a particular matter can be difficult. It is not obvious what changes to meeting formats may be appropriate or possible, but both their functions and formats should be ardently reevaluated.

If the lack of discussion was limited to conferences, perhaps that would be tolerable. Most of the workshops I have recently attended, however, have had similar formats. Even when panel discussions were planned, some have turned into series of miniseminars. It therefore appears to me that our community is actually tending to discount the worth of public discourse.

4. Proposal reviews

My personal experience regarding the proposal review process is very limited. I suspect, however, that this process is not working as intended either. Several of my observations about the paper review process apply here as well.

Both proposals and papers are reviewed by a small number of people. My experience with reviews of submitted manuscripts is that about one-third of the formal reviews are really useful (after having passed through several friendly reviews already). This may reflect the percentage of reviewers who are both sufficiently knowledgeable and willing to undertake a careful review. A similar fraction may apply to proposal reviews. Unlike manuscript reviews that may be published within a year, however, the public may not have an opportunity to comment on some proposed work for a few years, until after it has been funded, conducted, and published.

Several proposals I have reviewed during the last several years have been very terse. In fact, I often could not ascertain whether the investigators understood their problems well enough to undertake them. As the individual award funds decrease and the percentage of accepted proposals decreases, even without more scientists entering the field, the number of proposals will tend to increase. As the number of distinct program announcements increases, the total number of proposals also tends to increase. Investigators are writing more proposals and more proposals are being reviewed. There is little time to either compose or review them all well.

A study of the National Science Foundation proposal review process was conducted 18 years ago (Cole and Cole 1981). Two independent sets of reviewers were used for selected proposals in some research fields. Although no significant biases were revealed, it showed that a large percentage of funding decisions could have been reversed by using different sets of reviewers. This was true even for proposals whose averaged rank was in the "very good" category according to one set of reviewers. So a considerable amount of luck was involved in getting a proposal accepted.

What makes me most suspicious of the proposal review process is the poor quality of too much of the work I see. In the subject areas I know best, I would say there are a handful of colleagues who know the subject well enough to distinguish a really good plan from a well-presented but fundamentally worthless one. The chance of even one of them being part of the review process is not great. Even then, they may be reluctant to level appropriate criticism or be unable to persuade their fellow reviewers. Yet, in some cases, this will be the only real review of the work that occurs.

5. Summary

Do I learn something from most articles or presentations? From maybe half, my answer is no! Rather I am often left confused and wondering, "How do these investigators obtain such wonderful results when they make such fundamental errors or grossly poor assumptions? Can it be because their work is restricted to one case? Is it because of what they compare against? Is it the measures of success they use? Or, is it something

fundamental that I do not understand or about which I am mistaken?" Even when I have been motivated to spend many months investigating the problem further myself, I can rarely fully explain their results because I cannot actually duplicate their work.

I have discussed these observations with many fellow scientists. Some quite agree. Others say some degree of poor work has always been and will always be present. This is likely true, except it also seems that too high a proportion of work falls in the poor category, to the point that the present situation should be deemed unacceptable. The situation may have been as bad in the past, but then there were fewer papers, presentations, and proposals; a greater percentage of published comments; and significant public discussion at conferences and workshops. Real changes seem to have occurred during the past 20 years. There will always be some poor work, especially if enough people do not speak up and work to change the present situation, even if that means their own work will undergo greater scrutiny.

6. Recommendations

My paramount recommendation is that our community acknowledges that a major problem in fact exists and requires ardent attention. Unless this is acknowledged, the community will likely not even consider significant changes. I suspect that too many scientists, especially those with the authority to demand changes, will prefer the status quo. Even if this is true, however, a minority can rededicate themselves to critical review and accountability.

Some changes to consider are the following.

1) Do not duplicate conference formats at workshops, unless it is necessary for some specific reason. Instead, allow ample time for questions and discussion by limiting the numbers of formal presentations if necessary, *because the discussions are too important.* The public dialogue must not be limited to two–five minutes of quick questions followed by equally quick answers. In particular, follow-up questions and the joining in by others must be permitted. In fact, it should be strongly encouraged.
2) Ensure that all work gets *adequate public review* at some stage. It must be a public review to increase the likelihood of persons being present who are both knowledgeable and willing to ask probing questions. Design a format that may encourage real criticism.
3) When scientific controversies develop and become ongoing, workshops should be organized and funded to clarify the issues and publicize the debates. Scientific managers should demand this, and it should be demanded of them.
4) Encourage publication of scientific criticisms. It takes courage and hard work to write a good scientific criticism, and this is in part what science is about. Such works should not be termed "negative," but instead should be rewarded if they are sound.
5) An electronic journal should be established allowing the community to attach dialogues to papers, in the forms of questions, answers, comments, and replies. The often long and sometimes contentious anonymous review process can be augmented by a more efficient option of allowing a paper to be submitted along with reviews requested by the author rather than editor. The reviews would appear as attachments to the paper, and the reviewers identified so that they too become accountable. If an author cannot obtain such reviews, the former process would still be available.
6) A public proposal review process should be tested for some small programs. This can be done after the number of submitted proposals has already been reduced, for example, on the basis of letters of intent. The remaining investigators can then be invited to present their proposals to each other and additional invited experts. Not only is the review then conducted by many experts, but there is an opportunity to rebut criticisms, perhaps with the help of others, and there is a strong motivation to offer sound criticism regarding others works. This can also help produce a more coherent program as investigators are encouraged to modify their plans and explain their fit with other proposals prior to any funding commitments being offered.

Are more radical changes required? I used to think not, but the more I see a lack of realization of the problem and a neglect to make even minor changes, I suspect any real changes may have to be radical. They must be radical enough to produce real effects, not cosmetic ones. They will likely be unpalatable for some because it means they will have to pay more attention to important details in their work and thereby reduce their rates of apparent production.

No individual should attempt to impose solutions on the community without extensive debates on these matters, so I prefer not to offer further suggestions

here. All the consequences of possible changes cannot be easily anticipated. A community dialogue on these issues is required.

I would not have written this letter if I thought significant and appropriate changes were impossible. In fact, I think it would be easy once a consensus for their need was established. There is more I can say about the present situation, specifically why I think the situation is worse than I present here and speculations as to why that may be so, but I will leave that for future letters.

References

Cole, J. R., and S. Cole, 1981: *Peer Review in the National Science Foundation: Phase Two of a Study*. National Academy Press, 106 pp.

Ronald M. Errico
National Center for Atmospheric Research
Boulder, Colorado