

10. Experimental Design

- *essential components*
- *exploratory vs confirmatory experiments*
- *some models*
- *pitfalls*

10. Experimental Design

a. Introduction

Too often, experiments in atmospheric science are conducted without careful experimental design, and experimental design usually receives little formal attention in graduate programs. It is difficult to conduct controlled experiments in atmospheric science, and as a result many field projects produce inconclusive results, only suggesting new interpretations but seldom confirming them. The aim of this chapter is to discuss some aspects of a more formal approach to experimentation. This approach provides a valuable model even in those cases where measuring systems are inadequate or weather systems are too unpredictable to permit full use of these approaches.

A distinction is sometimes made between *experimental research*, involving use of these principles, and *observational research*, characterized by collecting a comprehensive set of measurements and then using them in exploratory analyses to learn about new phenomena. This does not seem to be a very useful distinction, because most experiments are strengthened by attention to these principles, so far as possible, and only differ in the degree to which hypotheses, critical measurements, statistical tests, etc., can be specified in advance.

b. Components in experimental design

An experiment is usually undertaken to study some general topic, such as how ice forms in clouds, how rainbands are formed, how entrainment affects droplet size distributions, how precipitation forms, or what conditions are required for contrail formation. The scientific objectives often include discriminating among a set of specific hypotheses, representing various alternative explanations. The goal in experimental design is usually to find observable consequences that distinguish among the hypotheses, and then collect measurements that can differentiate among those possibilities (or perhaps invalidate them all). This is always an interactive process: The statement of hypotheses must reflect the consequences of the physical processes in the particular application selected for observation, and the observations must be feasible. The elements in experimental design, although often presented serially, are almost always developed iteratively as compromises between what is possible and what would be decisive.

Those elements include:

- *A set of hypotheses.* If the hypotheses can be stated in very specific terms, the experiment often can be designed to provide critical and convincing tests that distinguish among them. For example, the hypotheses might be that various specific nucleation mechanisms will be responsible for the formation of the first ice in a particular cloud. These possibilities may have testable consequences that can be detected, and the experiment can be designed to ensure that the hypotheses can be differentiated. In another case, the hypotheses might specify the various possibilities for the dominant dynamical process leading to the formation of a rainband. It is of course never possible to prove a hypothesis, only to obtain evidence either consistent or inconsistent with the hypothesis, so the set of hypotheses for an experiment should include conventional explanations as well as new and more controversial possibilities. Often, a new hypothesis will arise during the exploratory analysis of data, but the experiment will be more convincing (and probably better designed) when the hypotheses have been stated explicitly in the experimental design.

- *Experimental tests.* When the hypotheses lead to different results, key features can be selected that would serve as tests. In the example of ice formation, the consequences of a particular nucleation mode might be that ice formation is governed by temperature or supersaturation or collision with aerosol particles; each of these possibilities leads to different ice formation in different parts of a cloud, and so has observable consequences. The selection of appropriate experimental tests is the key aspect of experimental design, requiring understanding of practical as well as scientific issues. One must select experimental conditions that occur with appropriate frequency, that can be recognized and probed with available instrumentation, and that provide good tests of the hypotheses.

- *Measurement strategies.* Many consequences that can be hypothesized are outside the capabilities of current measuring systems, so meteorological experiments must consider if the measurement strategy is practical and must identify the instruments needed to perform the experiment.

- *Analysis strategies.* An often-neglected component in experimental design is consideration of the analysis approach. For example, what sample size will be needed to draw conclusions of statistical significance? What tests will be applied to the data to accept or reject hypotheses? Experiments are strengthened when these can be specified in advance and considered in the experimental design.

c. *Exploratory vs confirmatory experiments*

Most experimental work in atmospheric science is exploratory in nature, and formal confirmatory experiments are seldom conducted. (An exception is in weather modification, where some confirmatory experiments have been attempted.) Confirmatory experiments are those that replicate an earlier experiment with specific statements of hypotheses, tests, measurement and analysis strategies, and expected outcome. Such experiments are especially important when the results of an exploratory experiment provide marginal statistical support for a hypothesis that arose during the analysis of the data. A confirmatory experiment usually tests a single hypothesis (e.g., that cloud seeding increases precipitation) vs the null hypothesis (e.g., that no increase in precipitation results from seeding) under fully specified experimental conditions.

An important reason for confirmatory experiments is that exploratory analyses risk the danger of *multiplicity*, the increased likelihood that some effects will appear to meet tests of statistical significance simply because many possibilities are considered. If 20 classifications of data are considered, on the average one will appear to meet a test of being expected only 5% of the time, so there is a danger that apparently significant results can result by chance when the possibilities are not limited in advance. A value of confirmatory experiments is that they minimize this danger by testing a specific hypothesis stated in advance.

Another argument for confirmatory experiments is that it is often difficult to know the underlying probability function governing meteorological events, which are often highly correlated in time and hence difficult to use with some approaches to statistical inference. Confirmatory experiments do not eliminate this problem but can reduce it when used with different experimental populations.

In truly exploratory experiments, where a large number of possible relationships are considered in the data, there is often value in reserving a portion of the data for confirmatory use before undertaking the analysis. Once trends are identified in a portion of the data, the remainder can be used to test for these same trends in an independent dataset.

d. *Some model experimental designs*

Studies of weather modification provide models because such studies have by necessity confronted some of the key problems facing meteorological experimentation, including natural variability, correlations of weather events in time and space, and biases of experimenters. Some features of experimental design used in weather modification are discussed in this section.

1). Target-control designs

One way to reduce the problem of high natural variability is to find a *covariate*, a quantity that is correlated with the quantity of interest so that it can be used to reduce the expected variance in that quantity. For example, if the rainfall in a particular area is correlated with some other independent quantity, like instability or low-level moisture, then knowledge of that other quantity can be used to predict the desired rainfall to some accuracy that is better than the climatological average. Such a covariate can make it possible to detect smaller effects of seeding on rainfall. (The same principle applies to other observed weather phenomena and to causes other than seeding.)

One covariate often used is the rainfall in a nearby area. If the nearby area is unaffected by seeding but historically shows a high correlation with the rainfall in the target area, then that rainfall can be used to predict the rainfall expected in the target area in the absence of seeding. For example, a linear regression model of the relationship between target and control precipitation might indicate a correlation coefficient of r between target and control precipitation, thereby reducing the variance about the regression prediction by a factor of $1 - r^2$.

There are serious dangers in use of target-control designs. Persistent weather patterns can produce short-term departures from historical relationships, so the linear regression model (applicable to randomly occurring events) can lead to serious errors when applied to such correlated sequences. Another flaw is that meteorological phenomena seldom occur with Gaussian distributions, but more commonly have more outlying events (e.g., with high precipitation) than a Gaussian distribution. Because of this, tests that rely on Gaussian distributions for their justification are usually not applicable. (Sometimes this problem can be alleviated by working with data transformed to new variables, such as square roots or cube roots of rainfall amounts, selected to give distributions closer to Gaussian in character, but this does not help the problems introduced by correlated sequences.)

2). Randomized target design

Randomization of experimental conditions can be of great value in determining if treatment affects the outcome, so randomization is used frequently in conjunction with cloud seeding experiments. Because cloud seeding is not an emphasis in this course, randomization will not be emphasized here, except to note that it can be used with any of the experimental designs. For example, if a target-control ratio for randomly seeded clouds is compared to a similar ratio for the unseeded clouds, the result is less affected by persistent correlations than is the target-control ratio when all clouds are seeded. An important variation of these designs is the randomized crossover design, in which one of two areas is seeded randomly and the other used as a control area.

3). Non-parametric tests and rerandomization

Because the distribution of natural events is often far from Gaussian and may not be known well, tests that do not rely on the nature of the probability distribution are needed to assess significance levels in many experiments using meteorological data. If it is found, for example, that rainfall in seeded cases averages 1.2 times as much as in unseeded cases, the significance level of this result can be assessed by considering a large set of simulated experiments for which the rainfall amounts are retained but the seeding decision is reassigned randomly. If there were no real seeding effect, then the fraction of these simulated experiments producing apparent seeding increases by factors of 1.2 or more can be used to determine the confidence limit to be associated with the result. The same approach can be used with other characteristics of the results. A common statistic used in this way is the sum of the ranks associated with seeded storms, for which the test is evaluated using a Wilcoxon test.

“Rerandomization” can be used even when there is no real randomization. For example, one way to test for the significance of an apparent correlation (e.g., between lifetime of cumulus clouds and dewpoint depression in the environmental air) is to rerandomize the values of one of the variables and recalculate the correlation coefficient with the

rerandomized variables, thus determining the probability that the measured correlation coefficient would arise by chance. This avoids the assumption of Gaussian distributions inherent in standard estimates of the errors in correlation coefficients. This can be a valuable technique when it is suspected that the distributions are not Gaussian.

A danger still present with rerandomization is that presented by correlated sequences in the data. If correlated sequences are rerandomized, too great a significance level may be attached to the result.

e. Some dangers in experimental design

1). Unrecognized causes

A correlation between two sets of observations does not necessarily indicate a causative relationship between the two measured characteristics. It is often the case that X and Y are correlated because both are related to Z , and neither X nor Y influence the other. For example, rainfall on a given day is highly correlated with rainfall on the preceding day, but this does not reflect a causative relationship so much as a tendency for weather patterns to persist for several days. When designing experiments that search for causative relationships, one must consider alternate relationships that will produce correlations like this one.

2). Climatology

Planning for expected weather events is often the most difficult part of designing field experiments. The success of field experiments often depends on encountering suitable weather, and those weather events may occur irregularly. Often suitable climatological information is not available because the experiment depends on detailed characteristics of the weather that are not documented routinely. Experience in the area where the experiment is planned is invaluable.

Extensive climatological data are available, from sources that include the National Climatic Data Center, the archives at the National Center for Atmospheric Research, and from many state climatologist offices. Useful contacts for these data are: at NCDC, National Climatic Data Center, Federal Building, Asheville, NC 28801-2733 (704-271-4800; FAX 704-271-4876; e-mail orders@ncdc.noaa.gov); at NCAR, Data Support Section, National Center for Atmospheric Research, P. O. Box 3000, Boulder CO 80307-3000 (e-mail datahelp@ncar.ucar.edu; World-Wide Web address <http://www.ucar.edu/dss/index/html>).

3). Type-I and Type-II errors

Two types of errors can occur in a well-formulated experiment consisting of a specified test of a hypothesis. A type-I error occurs if the hypothesis fails the test and is rejected even though it is true. A type-II error occurs if the hypothesis is not rejected even though it is false. Both occur because of statistical fluctuations present in sets of observations, but it should be possible to estimate the probability that this has occurred. For example, if a result lies more than two standard deviations from a prediction, we can reject the hypothesis that the prediction is correct and expect to be wrong only about one in 20 times *in a well-designed experiment*.

Some experiments, particularly those associated with weather modification, are best formulated in terms of the “null” hypothesis, the hypothesis that the treatment has no effect. For example, in a cloud-seeding experiment the null hypothesis may be that release of seeding material has no effect on the precipitation. To test this hypothesis, we determine if the precipitation falling from seeded clouds differs significantly from that falling from unseeded clouds. If so, one “rejects the null hypothesis” — i.e., concludes that precipitation differs in seeded vs unseeded clouds. (Note that this still does not imply that seeding caused the difference; the difference may arise because the seeded clouds were naturally more vigorous, or because of some other unrecognized cause.) In this case, a type-I error occurs if we reject the null hypothesis, and conclude there is a significant difference, when in fact there is none.

The experiment needs to be designed so that the expected result provides a definitive test. For example, if the result of a seeding experiment is that the null hypothesis is accepted (i.e., there was no significant difference between seeded and unseeded cases), this may only reflect an inadequate sample size or too small an effect to detect with the selected experimental design. Numerical experiments and statistical simulations can help avoid poor experimental designs. For example, one might want to test how well current formulations of collision efficiencies predict the rates of coalescence of water droplets leading to rain. Numerical experiments ahead of the field experiment can help determine the accuracy needed in measurement of such relevant factors as updraft speed, liquid water content, droplet size distribution, radar reflectivity, cloud condensation nucleus population, etc. These experiments can help avoid the all-too-common inconclusive experiment, one producing the primary conclusion that a more careful experiment is needed.

4). Effects of natural variability

Natural variability is the bane of most field experiments. Events in the atmosphere are usually not characterized well by Gaussian or other standard distributions, because extreme events occur more often than would be expected from the mean. Furthermore, the extreme events often dominate results like precipitation amounts or property damage. “Log-normal” distributions, in which events are distributed according to Gaussian distributions in the logarithm of a variable, often provide better representations of events in the atmosphere, but still are not reliable guides in most cases. When experiments are undertaken that rely on statistical comparisons, one must always consider the role of natural variability in such comparisons. Rerandomization, discussed in section d, is usually the only reliable way to account for the effects of natural variability.

A good example of the influence of natural variability is the measurement of the vertical flux of humidity in the atmospheric boundary layer. The flux of water is $\overline{\rho w}$ where ρ is the density of water vapor, w is the vertical wind, and the bar denote the average over a region in the boundary layer. To estimate the flux using measurements from a research aircraft, one can use

$$\overline{\rho w} \approx \frac{1}{N} \sum_i (\rho_i w_i) \quad (10.1)$$

where the summation includes N measurements that span a region in the boundary layer. There are several sources of uncertainty associated with such an estimate of the flux:

- The instruments used will have associated measurement uncertainties, so there are uncertainties in the individual measurements ρ_i and w_i .

- There is an uncertainty associated with the average obtained from N such measurements. This can be determined using the error-propagation methods of Chapter 2.
- For any finite sample from the boundary layer, there is uncertainty associated with using this sample to represent characteristics of the entire boundary layer.

The third source of uncertainty is often the dominant one. Any particular flight segment samples only one of many possible sequences that could be encountered in the boundary layer, and so there is uncertainty associated with using that sequence as a representation of the entire boundary layer. (Cf. Lenschow and Stankov 1986 for further discussion of this particular problem.) This is a pervasive problem in using a set of observations to represent extended fields. This type of problem is particularly critical in studies of the effects of small-scale processes on global climate, because the influences of ensembles of small-scale events are particularly difficult to determine in ways that represent the global influences of those events.

SOURCES AND FURTHER READING

- Anderson, V. L., and R. A. McLean, 1974: Design of Experiments. Marcel Dekker, Inc., New York, 418 pp.
- Dennis, A., 1980: Weather Modification by Cloud Seeding. Academic Press, New York, 267 pp.
- Murphy, A. H., and R. W. Katz, 1985: Probability, Statistics, and Decision Making in the Atmospheric Sciences. Westview Press, Boulder, Colorado, 545 pp.

page intentionally left blank