Reprint from

MARINE MESOCOSMS

Biological and Chemical Research in Experimental Ecosystems

Edited by George D. Grice and Michael R. Reeve

Springer-Verlag
New York Heidelberg Berlin

Contents of MARINE MESOCOSMS

16

Replication in Controlled Marine Systems: Presenting the Evidence

Woollcott Smith, Victoria R. Gibson, and J. Frederick Grassle

Introduction

The basic reason for replicating experiments is to answer the question: If a competent experimenter were to reconstruct this experiment with approximately the same experimental manipulations and environmental conditions, what range of results would he observe? Attempting to answer this question for large-scale experiments faces us with two problems not encountered in small-scale laboratory experiments: (1) Replicate experiments are performed under different environmental conditions, and (2) the cost and size of an experiment make large numbers of replicate experiments impractical. This note discusses these two problems in general and then considers the evidence for replicability in two large-scale marine experiments: the MERL No. 2 fuel oil experiments at the Marine Ecosystems Research Laboratory, Narragansett, R.I., U.S.A., and the CEPEX mercury pollution experiments (Controlled Environmental Pollution Experiment, Saanich Inlet, British Columbia, Canada).

Woollcott Smith and J. Frederick Grassle, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, U.S.A.

Victoria R. Gibson, Division of Biomedical Communications, University of Texas Medical Branch, Galveston, Texas 77550, U.S.A.

Comparing Treatment Responses in Fluctuating Environments

In general, our hypothetical competent experimenter will not work with the same community or system each time he repeats a large-scale marine experiment. Although he is working in a similar (e.g., temperate estuarine) environment, its physical, chemical, and biological properties will vary from experiment to experiment in many important details. Measures of treatment effect must, therefore, be independent of special characteristics of a particular community or set of environmental conditions. To achieve this independence one chooses as the measure of effect a general property of a marine ecosystem (e.g., primary production or species composition). The general property is then compared to a control or untreated system. In the language of statistics, the measurements in the control are used as covariates to remove the fluctuations in the experimental systems that affect both the control and experimental systems. With these variations removed, the effect of the experimental treatment should become more apparent.

In measuring a general property of a system, detailed physical and chemical measurements and species identification serve two purposes: They can be combined to estimate the selected general property of the system, such as secondary production, species diversity, or similarity index; and they can be used as covariates to remove variation due to uncontrolled environmental fluctuations. We have now defined in a general way how one would carry out a *single* paired experiment: (1) Our competent experimenter constructs at least two identical systems, control and experimental. (2) The treatment response is found by comparing control and experimental systems using a general quantitative measure of the difference between the two systems.

Assessing Replicability Using a Small Number of Experimental Units

We must now face the second major problem: How to predict the variability one might find if another competent experimenter were to attempt to repeat the experiment. A direct approach would evaluate the differences among responses in a number of single experiments performed separately by different experimenters. This procedure, widely used in clinical trials for experimental drugs, is too time consuming and impractical for most scientific work, so less direct evidence is used. One assumes that the conditions and variability within a single laboratory are equivalent to the variability across laboratories, so that the variability between replicate experiments conducted by a single experimenter can be used.

In many experimental situations (clinical medicine, sociology, or psychology) a single experimenter is limited to a small number of experimental units, due to their high cost, the labor-intensive nature of the work, or the unavailability of suitable units for comparison. For example, in evaluating the effects of government regulatory policy on a society (Box and Tiao 1975, Huitema 1979), the

experimental unit is the entire country or governmental subunit. The limited number of experimental units is also a problem inherent in large-scale marine experiments. This problem can be partially overcome by careful experimental design; for example, one could start with controls related very closely to the experimental units (see the description of the MERL and CEPEX experiments by Grice and Reeve, Chapter 1).

Another important difference between the ecosystem experiment and simpler experimental designs is ability to vary the experimental treatment of the ecosystem over time. Such time series experiments have been the subject of considerable research to determine the best design and analysis. Glass et al.'s (1975) book introduces time series experiments very well. Using similar techniques, Mulholland and Gowdy (1977) investigated the design of ecosystem experiments. Although the technical details can become rather cumbersome, the basic principles of design and analysis are straightforward and have been employed to some extent in large marine experiments.

In time series experiments one can conduct a series of experiments within a single experimental unit. The simplest and, for ecosystem experiments, perhaps the only practical form is the "ABA" experiment, in which control condition A is followed by experimental condition B, followed by a control period A. Such studies of the effects of chemical pollutants in large systems are further complicated: The treatments cannot be turned on and off at will, because the experimenter has little control over the processes affecting decay of the pollutants in the systems. For these systems, the experiment consists of a continuous buildup and decay of the chemical treatment. The strongest evidence for a causal effect in an ABA experiment is a response that increases and then decays in the same manner as the treatment. If the experimental and control systems converge again after the treatment stops, then there is strong evidence that the observed effect is not merely due to random divergence of the two systems over time.

Examples: Evidence from MERL and CEPEX

Let us consider how one might address the question of replicability in the MERL and CEPEX experiments. We will examine the response of the zooplankton community to the addition of mercury in CEPEX experiments carried out during the summers of 1976 and 1978 (Grice et al. 1977; Reeve et al., Chapter 28), and the response of the benthic community to No. 2 fuel oil in the MERL experiments during 1978 (Grassle et al. 1980). Data from these experiments are used to investigate two questions. First, within a single experimental unit, how well does the response time series follow the treatment time series? Second, how well do experiments conducted within different units replicate?

We have chosen dissimilarity in species composition (information overlap; Horn 1966) between control and experimental systems as the general measure of experimental effect. This dissimilarity is 0 if the relative proportion of each taxonomic group in the two systems is identical, and 1 if the two systems have no species in common. Our experience (Grassle and Smith, 1976, Smith et al. 1979)

indicates that species composition as measured by a dissimilarity index is a sensitive indicator of pollutant effects. One might also use other general measures of system response, such as primary or secondary production or species diversity.

The MERL Experiment

The MERL experiment was conducted in nine tanks (three experimental and six control). The unbalanced design was due to the need to reserve three additional control tanks for experimental treatments in the second year. The benthic community was sampled monthly beginning in November 1977 (details in Grassle et al. 1980). Twice weekly doses of oil were added to the experimental tanks between 6 March 1978 and 6 July 1978; in this period a total of 40.25 grams of oil was added to each experimental tank (Gearing et al. 1980). Individuals in samples were identified by species, but for computing dissimilarity indices the species were grouped into five major taxa: amphipods, bivalves, gastropods, polychaetes, and miscellaneous. Because of the highly variable settling rate of the polychaete larvae, particularly *Mediomastus ambiseta,* this taxon was removed from the analysis.

We used the following procedure to compute the dissimilarity between the experimental and control tanks. For a single sampling period we calculated the proportion of each taxon per tank. This proportion for the six controls was averaged to form a mean control. Next we calculated the dissimilarity between the mean control and the three individual experimental tanks (Figure 16-1). This is the response time series.

We also calculated the variability within the six control systems (Figure 16-1): The range of the dissimilarity between each control and the mean of the other five controls is plotted. The treatment time series, total saturated hydrocarbons in the sediments, is given in Gearing et al. (1980), and is replotted for tank 7 at the bottom of Figure 16-1. Hydrocarbon time series in the other two experimental tanks were similar.

The data in Figure 16-1 are smoothed according to the three-point median method (described in McNeil 1977), in which the smoothed dissimilarity estimate is the median of the three adjacent time points. This removes large fluctuations in the dissimilarity index due to isolated larval-settling events. This high frequency noise due to rapid larval events has been noted by Hall et al. (1970). Smoothing data over time is one way to deal with this problem. Another solution would be composite samples over time (Rhode 1976); the sampling procedure itself would smooth the result.

Figure 16-1 contains the information needed to assess the effect of the treatment time series on the response series. The treatment approximately follows the ABA design. The slight dissimilarities in February and March indicate that the goal of the design—obtaining nearly identical initial conditions in each tank—was achieved. All three experimental tanks' responses follow the treatment time series, with some indication of recovery in the October, November, and December sampling periods. Because of the time series nature of the data, a full statistical analysis would be complex. However, for each time point from May through

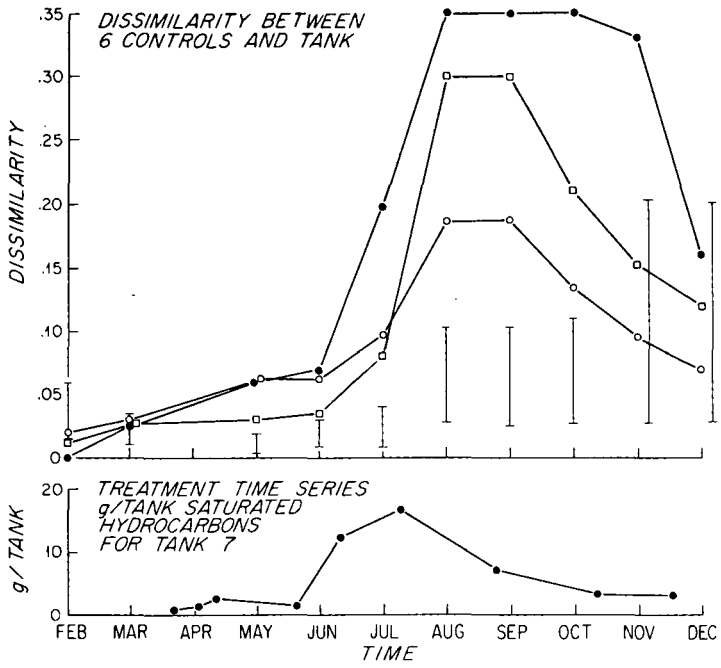

Figure 16-1. Response of benthic community to saturated hydrocarbons in the 1978 MERL experiment (Grassle 1980). Top: Horn's (1966) measure of dissimilarity between tanks 2, 7, and 9 and the mean of the six control tanks. Solid circle, open circle, and open square denote tanks 2, 7, and 9, respectively. Vertical bars give the range of dissimilarity when each control is compared with the remaining five controls. Bottom: treatment time series saturated hydrocarbons in tank 7 [after Gearing et al. 1980].

October, the dissimilarities for the experimental tanks differ significantly from the controls ($P < 0.01$ level, nonparametric Mann-Whitney U test).

The data presented (Figure 16-1) do not indicate whether or not the observed effect is general or particular to the 1978 experiment. Data from several years are needed. Similar results reported by Smith et al. (1979) from a 1977 experiment with similar but not identical conditions, indicate that the observed effect was not unique to the 1978 system.

The CEPEX Experiment

Experiments on low-level concentrations of mercury were conducted in the CEPEX enclosures in May 1976 (Grice et al. 1977) and in the late summer of 1978. The community dissimilarity measure described for the MERL experiment can also measure the effect of the mercury treatment on the zooplankton community in these enclosures.

In each experiment, a first enclosure was the control, a second enclosure was given a low dose of mercury (about 2 μg/liter), and the third enclosure received

a high dose (about 5 μg/liter). In the earlier experiment, titled Hg-III, the initial concentrations of mercury decayed as particulate matter sank to the bottom of the water column. In the 1978 experiment, Hg-V, concentrations were maintained by adding mercury throughout the duration of the experiment.

We calculated the dissimilarities between zooplankton communities, as in the MERL experiments, but used all 19 major zooplankton taxa (Figure 16-2). The Hg-III and Hg-V experiments, conducted 2 years apart and in different seasons, showed similar quantitative responses to both the high and low mercury concentrations (Figure 16-2). Together, these replicated experiments present convincing evidence of a general response of the zooplankton community to low levels of mercury pollution.

In the CEPEX experiments, pre- and postexperimental control periods were less well defined than in MERL experiments. In addition, because of the unstable

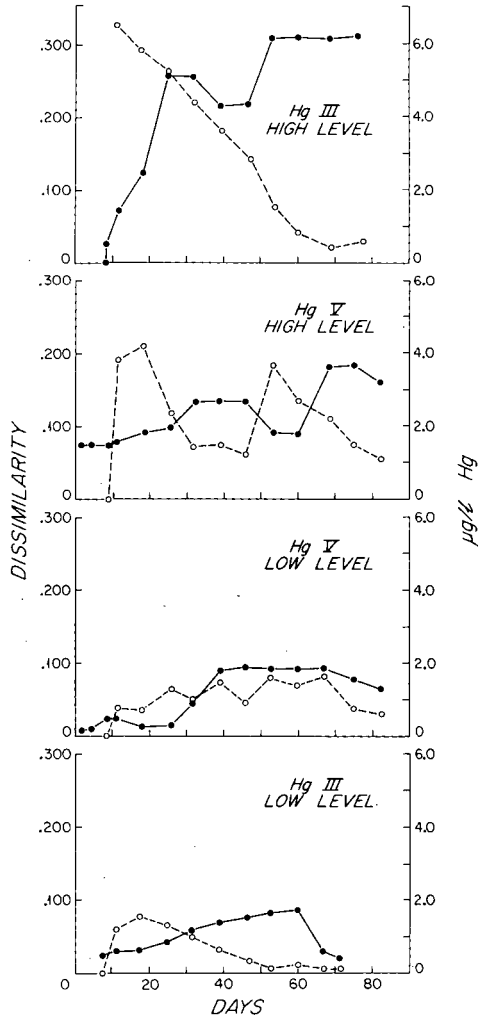

Figure 16-2. Response of the zooplankton community to mercury in two separate CEPEX experiments, HG-III and Hg-V, in May 1976 and August 1978. Horn's dissimilarity measure (*dashed line*) was calculated between the experimental enclosure and the control using 19 major zooplankton taxa (Grice et al. 1977). The treatment time series, the concentration of mercury, is indicated by the solid line.

nature of the enclosed zooplankton communities (see Steele 1979) it was less likely that dissimilarity would decrease in the postexperimental period, indicating recovery. There is, however, an indication that recovery did occur in the low-level experiments.

Discussion

Large enclosed marine ecosystems (MERL and CEPEX) were designed to solve the problems of measurement and sampling in natural marine systems and also to overcome some of the problems of scale associated with microcosm experiments in laboratory aquaria. In enclosed marine experiments the problems of designing statistical sampling can be investigated using classical statistical methods (Lawson and Grice 1977, Williams et al. 1977). However, the experimental design and data analysis for these experiments differ in many respects from rigidly controlled experiments in aquaria.

A carefully designed and analyzed laboratory experiment is exemplified in a study of a freshwater system (Brockway et al. 1979), consisting of 30 small 8-liter experimental units in which all important variables (light, temperature, etc.) were carefully controlled. In this kind of laboratory experiment there is less need for complex data analysis to remove uncontrolled independent sources of variation. Thus, the remaining variation can be studied simply by evaluating the variability among the many independent experimental units.

Field experiments also can have many experimental units; Hall et al. (1970) describe an experimental freshwater system with 18 ponds in the experiment. Although classical experimental designs can be used successfully in these experiments, attention must still be paid to additional variation due to uncontrolled physical variables such as light and temperature.

The experimental-design and data-analysis problems for large enclosed systems are similar to those encountered in many experimental sciences in going from rigidly controlled laboratory experiments to field experiments. In this context, both the MERL and CEPEX experiments discussed here meet the conditions necessary for replicability in field experiments. That is, after the variation due to uncontrolled natural fluctuations is removed, the time series responses in all experimental units are similar.

Summary

The statistical design and analysis of large enclosed marine experiments differs from small well-controlled laboratory experiments. Although there are usually only a few large enclosures in an experiment, information on the replicability of an experiment is contained in the time series response in each enclosure. This is particularly true when the experimental unit is subject to a control period, followed by an experimental period and then a return to control conditions.

Zooplankton response to mercury in the CEPEX experiment and the benthic community's response to No. 2 fuel oil in the MERL experiment were analyzed.

Community dissimilarity was used as the measure of difference between the experimental and control systems. In both experiments the time series response, i.e., the dissimilarity between control and experimental communities, followed a pattern similar to the concentration of pollutant in the experimental systems.

Acknowledgments

L. S. Brown-Leger, N. J. Copley, and J. G. Smith Derby provided able technical assistance. We thank J. P. Grassle, G. D. Grice, and T. J. Lawson for valuable advice and assistance. The MERL studies were supported by EPA Grant 806072020 to the University of Rhode Island, and the CEPEX research was supported by the office of IDOE of the National Science Foundation, Grant OCE77-27446. Support for W. Smith was provided by NOAA-Sea Grant Program at the Woods Hole Oceanographic Institution, 04-8-M01-149. This is contribution No. 4789 from the Woods Hole Oceanographic Institution.

References

Box, G. E. P., and G. C. Tiao. 1975. Intervention analysis with application to economic and environmental problems. *J. Am. Statist. Assoc.* 70: 70-79.

Brockway, D. L., J. Hill, IV, J. R. Mandsley, and R. R. Lassiter. 1979. Development, replicability and modeling of naturally derived microcosms. *Int. J. Environ. St.* 13: 149-158.

Gearing, S. N., R. J. Pruell, T. L. Wade, and S. G. Quinn. 1980. Partition of No. 2 fuel oil in controlled estuarine ecosystem. Sediments and suspended particulate matter. *Environ. Sci. Technol.* 14: 1129-1135.

Glass, V., V. L. Willson, and J. M. Gottman. 1975. *Design and analysis of time-series experiments.* Colorado Associated University Press, Boulder, Colorado.

Grassle, J. F., and W. Smith. 1976. A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. *Oecologia* 25: 13-22.

Grassle, J. F., J. P. Grassle, L. S. Brown-Leger, N. J. Copley, and J. G. Smith Derby. 1980. *Quantitative studies on benthic communities in experimental ecosystems.* Report for year 2 of EPA Grant 806072020. The Marine Ecosystem Research Laboratory, University of Rhode Island, pp. 144-168.

Grice, G. D., M. R. Reeve, P. Koeller, and D. W. Menzel. 1977. The use of large volume (1300 m^3), transparent, enclosed sea-surface water columns in the study of stress on plankton ecosystems. *Helgol. Wiss. Meeresunters.* 30: 118-133.

Hall, J. D., W. E. Cooper, and E. E. Werner. 1970. An experimental approach to the production dynamics and structure of freshwater animal communities. *Limnol. Oceanogr.* 15: 838-928.

Horn, H. 1966. Measurement of overlap in comparative ecological studies. *Am. Nat.* 100: 419-424.

Huitema, B. E. 1979. An analysis of covariance Markov model for interrupted time series experiments. *Commun. Statist.-Theor. Meth.* A8: 789-797.

Lawson, T. J., and G. D. Grice. 1977. Zooplankton sampling variability: Controlled ecosystem pollution experiment. *Bull. Mar. Sci.* 27: 80-84.

McNeil, D. R. 1977. *Interactive data analysis.* John Wiley and Sons, New York.

Mulholland, R. J., and C. M. Gowdy. 1977. Theory and application of the measurement of structure and the determination of function for laboratory ecosystems. *J. Theoret. Biol.* 69: 321-340.

Rhode, C. A. 1976. Composite sampling. *Biometrics* 32: 273-282.

Smith, W., D. Kravitz, and J. F. Grassle. 1979. Confidence intervals for similarity measures using the two sample jackknife. In L. Orloci, C. R. Rao and W. M. Stiteler, Eds., *Multivariate methods in ecological work.* International Co-operative Publishing House, Fairland, Maryland, pp.253-262.

Steele, J. H. 1979. The uses of experimental ecosystems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 286: 583-595.

Williams, I. P., V. R. Gibson, and W. K. Smith. 1977. Horizontal distribution of pumped zooplankton during a controlled ecosystem pollution experiment: Implication for sampling strategy in large-volume water columns. *Mar. Sci. Comm.* 3: 239-253.