Observing-Systems Simulation Experiments: Past, Present, and Future

Charles P. Arnold, Jr.[1] and Clifford H. Dey[2]

Abstract

A summary of the history of observing-systems simulation experiments (OSSEs) is presented together with a description of current methodology, its capabilities and limitations, and considerations for the design of future experiments. These experiments are defined as a type of sensitivity study and are contrasted with real-data experiments otherwise known as observing-systems experiments (OSEs), data-impact, or data-denial experiments, which form a related type of sensitivity study. Simulation is presented as a means by which an a priori evaluation of proposed remote-sensing systems can be made.

1. Introduction

This paper reviews the history of simulation experiments designed specifically to evaluate the potential contribution of proposed observing systems to forecasts generated from numerical-weather-prediction models. The motivation for conducting such experiments has shifted from an emphasis on designing an optimum observing system during the planning for GARP to an evaluation of proposed space-based remote-sensing instruments anticipated for the next century. Among the concerns of planners are instrument costs, the benefits that may be realized and how they may be measured, and consistent user requirements. While cost-benefit analyses may be difficult or even impossible to conduct in this regard, it is at least possible to evaluate the benefits realized in the form of improved forecasts using simulation techniques. Additionally, a set of user requirements can also be evaluated in order to determine internal consistency between individual requirements.

Passive remote sensing of the earth's atmosphere from space-based platforms is now a mature technology. After nearly a quarter of a century, instrumentation has matured to the point that meteorologists are now considering various active instruments, such as lidar, for even greater accuracy in probing the atmosphere. Estimated costs of research, engineering design, and fabrication, however, have begun to reach heights not too distant from where the instruments themselves will orbit. The Department of Defense, the Department of Commerce, and NASA are under increasing pressure to hold these costs down. Whereas a passive microwave radiometer, which provides vertical temperature profiles, may cost the federal government three million dollars or so, it is likely that it may run as high as $500 million to develop and place into orbit a single prototype active lidar, which would measure atmospheric winds.

The anticipated costs of active remote-sensing instruments are not the only cause for concern. The impact of passive sounding systems on numerical forecasts is generally accepted to be strongly positive in the Southern Hemisphere. Furthermore, these data have been crucial to making medium-range forecasts a reality. However, the impact of remote-sounding systems on short-range forecasts over the data-rich Northern Hemisphere continental areas is less clear, and operational numerical-weather-prediction centers are still quite cautious about admitting these data to the objective analysis schemes in those areas. It seems prudent, therefore, that before advancing any replacement system, passive or active, we consider its potential contribution to forecast improvements.

The contradictory results obtained from the use of satellite sounder data may also be related to the manner in which new observing systems are specified. Requirements which contractors use for the engineering design of an observing system are usually stated in terms of performance specifications such as horizontal and vertical resolution, absolute accuracy, frequency of observation, etc., which are provided by the prospective user. It is generally not known whether there is an internal consistency among these requirements. In other words, it may be that the possible benefit realized from meeting one requirement, such as resolution, could be offset by the equally possible detrimental effects of meeting another, such as frequency of observation. Although simulation can be used to examine incremental improvements or degradation associated with satisfaction of one or more of these specifications, it has not, to our knowledge, been used either in the design of an individual instrument or in the integration of that instrument within an observing system. This question on the use of simulation to assess the impact of a variable's information content in designing observational systems was perhaps first raised by Smagorinsky et al. (1970). In their work the variables were meteorological ones, such as surface pressure and boundary-layer wind. There is no reason, however, why the methodology could not be extended to include the types of specifications mentioned above. A basic question, then, to be asked prior to placing a new instrument on a spacecraft should be, what effect will a change in one or more of the individual performance specifications have on a forecast generated by a numerical-weather-prediction model which will utilize the data from this instrument?

Our review begins with the various terminology and definitions associated with numerical simulation. Section 3 reviews the literature of the past 31 years related to observing-systems simulation experiments (OSSEs), and Section 4 summarizes the capabilities and limitations of simulation and from this perspective suggests a design for future experiments.

[1] Lt. Col., USAF, assigned to NESDIS.
[2] National Meteorological Center, Development Division.

2. Definitions and terminology

As with almost every subject today, there is an array of specialized terms associated with meteorological simulation. As used herein, simulation implies a process by which the true atmosphere or observations thereof are approximated by imperfect models and data. The model in this case is considered to be a complete assimilation/forecast system consisting of an analysis method, initialization technique, and numerical prediction model. An excellent description of these latter terms may be found in McPherson (1975).

Simulation studies or experiments are often used to infer the impact due to the loss or gain of simulated data on the forecast produced by a numerical prediction model. When real data are used for the same purpose, the experiments are referred to as observing-system experiments (OSEs), data-impact studies, or data-denial studies. Together with simulation experiments they are collectively called sensitivity studies. This paper will focus on simulation. The reader interested in data-impact studies is referred to the following abbreviated list of references (Ohring, 1979; Tracton et al., 1980; Tracton et al., 1981; Schlatter, 1981; Atlas et al., 1982; Koehler et al., 1983; Yu and McPherson, 1984; Barwell and Lorenc, 1985).

Simulation may be used to evaluate observing, analysis, or forecast systems, individually or in combination, thereby isolating the contribution of each to the analysis or forecast error. Observing systems include any proposed observational instrumentation (be it ground, air, or space based, passive, active, or in situ) for which the performance specifications are known. The latter must be known in order that realistic observational errors can be specified. An OSE may also be used to evaluate the effects of varying observing-station density. These have been referred to as observing-network density studies. Analysis and forecast simulation experiments are run in order to assess the effects changes to one or both will have on resultant analyses or forecasts. Although a few early papers concerning analysis experiments and network-density studies are mentioned, we will concentrate this review on observing-system simulation.

Part of creating simulated or fictitious observations, whether they be from real or proposed instrumentation, is the proper handling of their associated errors. Errors are of two types—systematic, otherwise known as mean or bias errors, and random errors. Frequently both are grouped into a single category, observational error (Alaka and Lewis, 1967). Random errors may result from internal instrument noise or external factors related mainly to how well the instrument samples the atmosphere. Random errors may also include a contribution resulting from the response of the instrument to scales of motion not resolvable by our numerical-weather-prediction models. The latter effect may account for a substantial portion of the random error. A discussion of these errors can be found in Wilcox and Sanders (1976) and Bruce et al. (1977). Both of these papers point out the need, particularly in the case of the random error, to partition this error into contributions arising from both the satellite and the rawinsonde. Instrument bias is simply the average difference between the instrument readings and a standard, which may be an internal calibration or a comparison with another instrument. For example, remote-sensing instruments used to measure vertical temperature profiles may have a bias of ± 0.5K and a random error of 1–2K root-mean-square (rms) when compared to rawinsonde.

The most-common method used to generate simulated observations is to first extract grid-point values from an extended run of a numerical predictio model. The complete outoutput from the model has been referred to in the literature as either the truth, nature, history, or reference atmosphere. We have chosen to use the latter, a term introduced by Alaka and Lewis (1967). For an assumed distribution of observations which would be available from a specified observation system, a representative set is extracted from the reference atmosphere, after properly interpolating grid-point values to the observation locations. Appropriate errors are then added to these simulated values to form the final set of simulated observations. Known biases can be added algebraically to each observation while the rms errors are added by first generating, with the help of a random-number generator, a population of errors which when added to the observations result in the desired rms error between the simulated observations and the standard. The final set of simulated observations can then be assimilated into a forecast model.

The success of a proposed observation system can be judged by how well the analyses, incorporating the new data, fit the reference atmosphere (RA) or how closely the forecasts based on the use of simulated data approximate the RA relative to corresponding forecasts made from the RA itself. The latter represents the best forecasts possible with a given model. When both the assimilation method (A/F) and the model used to generate the RA are one and the same, the experiment is referred to as an "identical twin." When both models are similar but not identical we have chosen the name "fraternal twin." Finally, if the RA consists of a series of analyses instead of a numerical-model output, we have chosen to call the experiment an analysis-series experiment. Figure 1 illustrates these three types of simulation experiments.

3. Historical review: 1954–1985

Newton (1954) was perhaps the first to suggest that variations of fictitious observations could be used to evaluate their effect on numerical forecasts. Given the impossibility of actually adding and removing observation stations until an optimum state could be reached, Newton suggested that an alternative would be to experiment with forecasts using analyses constructed partly from fictitious observations. Such experiments could be used to test the following: 1) where stations should be added on the fringes of present dense networks; 2) the effects of different possible interpretations of existing data; 3) the required accuracy of radiosonde; and 4) the network density required for satisfactory analysis.

As we shall see, the first, third, and fourth objectives remain as much a part of today's goals for OSSEs as they did 30 years ago.

Following the suggestions of Newton, Best (1955) investigated the differences between 500-mb barotropic forecasts resulting from different analyses of the same initial data. This study was a form of identical-twin simulation wherein the same model was used to generate forecasts initialized with two different analysis schemes. Bristor (1958) tackled New-

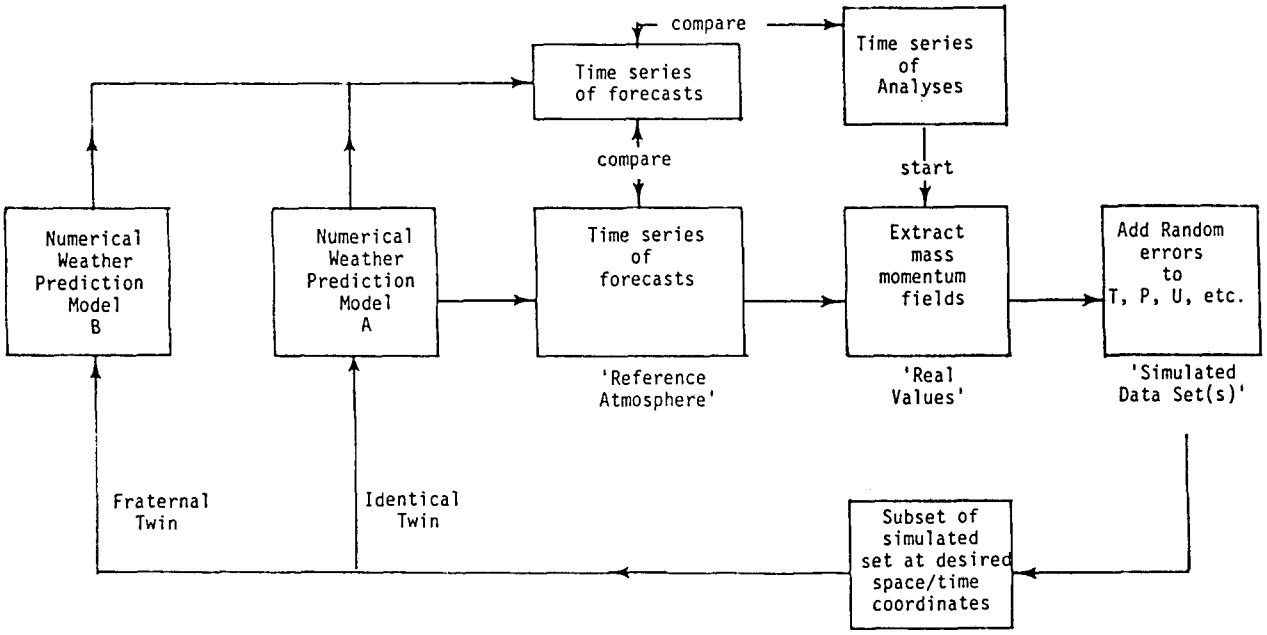

FIG. 1. Types of simulation experiments.

ton's question of required network density for satisfactory analysis by taking a hypothetically correct analysis and interpolating height and wind observations to simulate variation in network density. Bristor went so far as to add representative observational errors to these "perfect" observations to better simulate real data. Analyses created from each simulated network were then differenced from the control analysis to determine initial analysis error under conditions of varying station density. Forecasts computed from the differing analyses were also compared. This was an identical-twin experiment. Much of the methodology used in many of today's simulation studies was first developed in this paper. Jess (1959), who had not been aware of Bristor's work until the former had submitted his manuscript, developed an experiment very similar to that of Bristor's which, not surprisingly, showed broad general agreement. These four papers, often forgotten in the current literature, provided a foundation for subsequent research, initiated in the late 1960s, directed more towards OSEs, rather than network studies, or analysis-system experiments.

For the next 10 years there was little visible activity in the area of numerical simulation with the exception of Alaka and Lewis (1967). This paper, which begins with a rather nice discussion of analysis errors, reports on a network-density study. Contributing to the analysis errors are 1) the analysis method itself, 2) observational errors, both systematic and random, and 3) network density.

The latter contributes to analysis error through unresolved subgrid scale components which appear as noise (i.e., part of the random error) associated with larger-scale components. This effect the authors identify as aliasing. If large-scale features are to be forecast with higher accuracy, the aliasing portion of the observational error must be a small portion of the total power of the spectrum. They conclude that if a typical power spectrum for a given atmospheric field were known through its entire range, networks could be designed to satisfy this signal-to-noise issue. Unfortunately, as they point out, the short-wave end of the spectrum is not well known for most atmospheric fields.

Alaka and Lewis go on to relate network density to forecast accuracy by arguing that not only does the network density affect the magnitude of the initial analysis error, which contributes to forecast accuracy, but it affects the growth of the initial errors as well. The authors imply that error growth rate is directly proportional to network density, an assumption not borne out by earlier research (Bristor, 1958; Jess, 1959). Although not mentioned in their later work, (Gandin et al., 1967), these results showed agreement with both Bristor and Jess.

The Global Atmospheric Research Program (GARP) was initiated in 1967. One of its stated objectives was to determine acceptable compromise solutions to the data-requirement problem. In order to meet this objective, the U.S. Committee for GARP proposed "a national effort in computer simulation study of the predictive consequences of proposed observation systems to be known as 'Observing Systems Simulation Experiments'" (U.S. Committee for GARP, 1969).

Motivated by the GARP objective, Charney, Halem, and Jastrow (1969) designed two experiments for the purpose of determining 1) whether it would be possible to obtain the large-scale wind field from temperature alone; and 2) whether instantaneous low-level temperature and winds could be inferred from a continuous monitoring of their values at higher levels together with surface pressure. Their goal was to determine whether or not the implementation schedule of GARP could be advanced with satellite vertical temperature profiles which were expected to be available in the very near future.

Each experiment was of the identical-twin type, using a global circulation model (GCM) to predict winds and temperatures at 800 and 400 mb and pressure at sea level. The reference atmosphere was generated by integrating the model to 170 days. In their first experiment, a random

temperature error of 1°C was introduced at day 85 and the resulting circulation recalculated for 10 days. At day 95, rms "errors" were calculated for T, U, and V at both levels by comparing the forecast values obtained using the simulated temperature observations with the observed values in the RA. Beginning with the same day (day 95), the RA temperatures were inserted at intervals of one, six, 12, and 24 hours and the circulation recalculated for 60 days to day 155. It was shown that the greatest reduction in the rms errors resulted when the insertion of the "correct" values was made every 12 hours. Using this optimal insertion interval,[3] the calculations were again repeated with varying degrees of random error ranging from 1° to 0.25°C. As anticipated, the rms errors decreased in time and were proportional to the initial error.

Halem and Jastrow (1970) soon followed their initial work with a similar identical-twin experiment in order to evaluate the data requirements established for GARP. By adding various random errors to the initial conditions, forecasts were generated which were then compared to the RA values. The resulting rms errors in wind were compared to an arbitrary value of 6 $m \cdot s^{-1}$, defined as the limit of predictability. It was shown that when the initial errors are the GARP requirements of ± 3 $m \cdot s^{-1}$ rms error for the wind, ± 1°C rms error for the temperature, and ± 3-mb rms error for the surface pressure, the predictability limit was reached in three days.

Jastrow and Halem (1970) then asked whether the proposed GARP requirements were adequate or if a tightening of these requirements was necessary. Their results revealed that the GARP requirements were internally inconsistent. The error limits for the wind components and pressures were too large in comparison with the error limits set on the temperature requirement. Since it would be the latter that would be available from satellite, it was suggested that the error limits be reduced from 3 $m \cdot s^{-1}$ to 1.5 $m \cdot s^{-1}$ for wind observations and from 3 mb to 2 mb for pressure. This action would insure that forecasts generated from temperature data alone would not be degraded by using "poorer" quality wind and pressure observations and that, conversely, if pressure and wind observations were to be improved to the recommended levels, the range of accurate forecasts should be substantially extended. Such a tightening of requirements, although desirable, would have presented obvious design problems for the Global Observing System, since techniques for direct measurement of global winds with accuracy appreciably better than 3 $m \cdot s^{-1}$ were not available. Jastro and Halem's paper presented, for the first time, a methodology for evaluating the internal consistency of a set of meteorological requirements.

Williamson and Kasahara (1971) supported the Jastrow and Halem (1970) contention that GARP requirements were inconsistent. They showed, for example, that a 0.6°C rms temperature error (comparable to a maximum error of 1°C) leads to a 1.3 $m \cdot s^{-1}$ rms wind error when temperatures were updated. However, a 1.6 $m \cdot s^{-1}$ rms wind error (comparable to a maximum error of 3 $m \cdot s^{-1}$) leads to a 0.6°C rms temperature error when winds were updated.

[3] The European Centre for Medium Range Weather Forecasting has recently showed that 12-hour assimilation produced better results than 6-hour updating (Warburton, 1984). Similar results have been noted at the National Meteorological Center (McPherson, 1985).

Williamson and Kasahara (1971) were the first investigators to question the use of an identical-twin experiment and proposed instead an approach that we have chosen to call a fraternal-twin experiment. They noted two problems with the identical twin: 1) inaccurate treatment of physical processes, and 2) computational error in solving the model equations. To alleviate the first, they suggested using the most sophisticated model available to generate the RA and a less sophisticated model for the assimilation and forecast (A/F). To reduce the second, they suggested the RA be generated from an 2.5° grid while the A/F model be run on a 5° grid.

Gordon et al. (1972) performed an identical-twin experiment with their principal objective being whether wind observations would be necessary in the tropics if wind could be derived from existing mass field information. From assimilation runs using a nine-level GCM updated at 12-hour increments, they found that asymptotic rms wind errors in the equatorial tropics failed to meet GARP requirements for FGGE when only mass field data (surface pressure and temperatures at all levels) were used for updating. Their results also indicated that tropical wind data, assimilated at just two levels in the vertical, led to a significant decrease in rms wind error but not to the level of the GARP requirement. However, the insertion of wind data at all model levels reduced the wind errors to within acceptable limits. Among the limitations in this experiment were that it was an identical twin, that simulated observations were inserted at all grid points rather than at observation location, and that no errors were assumed for the surface pressure. An interesting suggestion made by the authors of this paper, but which to our knowledge has not been carried further, would be to investigate which scales of motion contain the largest errors.

Kasahara and Williamson (1972) introduced the term *model dependency* as a limitation to simulation studies. They questioned the interpretation of OSSE results, particularly when they were derived from an identical-twin methodology, in view of the work of Morel et al. (1971). The latter paper had discussed a "rejection phenomena" which might occur when data simulated with a different model or real data were used for updating experiments, which presumably would be absent and therefore not very simulative in an identical-twin experiment.

Kasahara (1972) again raised the question of model dependency relative to the identical-twin experiments by suggesting that the effects of "inherent" errors are common in both the model and the data, and hence the data would not be entirely foreign to the model, suggesting a greater compatibility and perhaps a resulting comparison that would be unrealistic. In addition to, or perhaps at least partly a result of, model dependency, he provides yet a second pitfall in OSSE, namely the difficulty one has in interpreting its results. Kasahara's paper is also notable because it modified the original GARP requirements for wind, temperature, and pressure to agree with the results of Jastrow and Halem (1970) and Williamson and Kasahara (1971).

Williamson (1973) also considered the effects of model dependency. He correctly identified the only source of forecast error in an identical-twin experiment as the model's *predictablity error growth,* which simply means that since the error growth rate of the model remains constant, the forecast error can only result from altering the initial conditions. Further-

more, since there is no additional accumulated error due to model imperfections in physics and numerics, the asymptotic error levels will be smaller and more optimistic.

Jastrow and Halem (1973) provide an excellent step-by-step description of how a typical simulation study proceeds. They also discuss the limitations of OSSEs. The most important limitation is identified as being a "compatibility issue," defined earlier by Kasahara (1972) as model dependency. However, Jastrow and Halem identify model dependency as a second major limitation in still a different context than Kasahara's. They concluded that whether one uses the identical-twin or the fraternal-twin approach, the physics of the simulated observations and the assimilation/forecast model are either identical or so close to being identical that this compatibility results in unrealistically small errors between the model forecast and the RA. On the other hand, in the real world, observations with "real world physics" tied to them are assimilated into a model with significantly different physics, such that the resulting incompatibility produces a greater difference between the real world and the model forecast. From this, they concluded that simulation studies on the hypothetical performance of an observing system always tend toward a more-favorable result than can be expected from the real observing system.

Jastrow and Halem (1973) regard model dependency as being solely a result of an identical-twin experiment. They explain that "the effect of minor defects in the physics of the model should cancel in forming this difference [between an assimilation with and without a set of simulated observations] of two circulations," but that "if the model has serious deficiencies the information on errors may be misleading." They therefore define model dependency as a sort of bias which, by virtue of the identical-twin method, is found in both RA and forecast and is such that some gradients may be reduced unrealistically.

By 1974 the Joint Organizing Committee for GARP specifically ruled out identical-twin experiments insofar as determining which of the possible special observing networks would be most effective during the First GARP Global Experiment (FGGE) (Lorenc, 1975). With this in mind, and building on the experience and recommendations of Williamson and Kasahara (1971), Lorenc performed an OSSE based on the several observing networks suggested by Bengtsson and Morel (1974). Therefore, rather than evaluating the contribution of a specific observing system, Lorenc considered the effect of an entire network of simulated observations from a variety of observing systems. The model used to generate the RA was the British Meteorological Office's five-level GCM. The same model was used for the A/F model with the grid increment increased by 50 percent. Improved versions of physical parameterizations were used in the RA while the original physics were retained in the A/F model.

Lorenc (1975) also included an analysis-system simulation experiment which considered three different analysis methods. Using the same observation network, model performance was evaluated using simple updating, which involved direct insertion of simulation observations at grid points, an optimum interpolation scheme, and optimum interpolation with the addition of climatological data. Climatology was weighted 20 percent and included in the background field prior to the assimilation of the simulated observations.

The first review of OSSE studies was conducted by Nitta (1975). He placed all previous work into one of two phases. Phase I included all identical-twin experiments and Phase II all fraternal-twin studies. Looking ahead to when the FGGE data sets would be available, he suggested that Phase III experiments should be based on an RA created from a time series of analyses taken from these data. From such an RA model, simulated observations could be generated and assimilated into an A/F model. The forecasts thus generated could be compared directly to the best possible analysis. Precisely why this approach should be an improvement over the fraternal-twin experiment was not made clear. Although it may seem that the best possible A/F model should provide the best assessment of the contribution of different observation systems, as first implied by Jastrow and Halem (1973), it must remain for future experiments to provide a clearer answer.

Following the FGGE period, there was little activity in OSSE work until the early 1980s. This was no doubt a direct result of the perceived decreased need to perform such experiments. Without a clear-cut objective, research in OSSE waned.

Among the few papers published in this period, there is one (Cane et al., 1981) that stands out. The objective of this paper was to assess the potential impact of space-based remote-sensing measurements of the marine surface wind field on numerical weather prediction. In view of earlier OSSE recommendations, it is difficult to understand why the authors chose to use an identical-twin approach. They point to the necessity of using a synthetic version of nature rather than a time series of real analyses because their goal was to study the effects of observational data that were not presently available. But this apparent obstacle could have easily been overcome by synthesizing observations from real analyses. They note that the errors and error growth rates they obtained were probably unrealistically low due to their use of the identical-twin method. We suspect that cost and/or availability of either a second model or real analyses was the most likely reason this experiment and others to follow were not designed differently.

Aside from these minor difficulties, Cane et al. (1981) is one of the few, if not the only, OSSE papers which addresses the question of statistical significance and sample independence. The authors suggest that five days between successive forecasts appears to be the minimum separation that allows the statistics to be independent. They did not, however, provide any support for this contention. Whereas Cane et al. compute a mean rms by averaging those from five independent forecasts together with their associated verifying analyses of the RA, others have apparently been satisfied that only one forecast would be sufficient since a vast number of individual grid-point values go into the rms computation. Atlas et al. (1981) are content with a two-day separation between 10 successive 72-hour forecasts. Exactly what length of time between successive forecasts is necessary to insure independence remains an open question in the conduct of OSSEs.

By mid 1980, the U.S. Air Force's meteorological satellite program was beginning to consider active laser-radar (lidar) methods for measuring the global wind field. Following a feasibility study conducted by NOAA's Wave Propagation Laboratory (WPL) and supported by the Air Force, it was

concluded that a simulation experiment should be performed. Simulated observational errors and performance characteristics of WPL's proposed *WINDSAT* instrument (Huffaker, 1978) would be used in the experiment. WPL asked the National Meteorological Center (NMC) to carry out the simulation.

In late 1980, a simulation project was established at NMC. Among its goals were 1) to develop an understanding of the data requirements for NMC's numerical analysis-prediction systems and 2) to develop a means of evaluating the potential usefulness of proposed new observation systems and changes for existing ones in meeting these requirements. An ambitious multi-year effort was built around the *WINDSAT* simulation experiment. However, because of the lack of adequate human and computational resources and financial support for *WINDSAT*-related research, it wasn't long before the project lost momentum.

Then, in early 1983, NMC joined with the Goddard Laboratory for Atmospheric Sciences (GLAS)[4] and the European Centre for Medium Range Weather Forecasts (ECMWF) in a new cooperative program for conducting and evaluating OSSEs. A workshop on the design of credible simulation experiments was held at NMC in February 1983. All participants agreed that although serious deficiences existed and careful interpretation was required, OSSEs offered the only available method for providing decision makers with information necessary to determine whether a proposed observing system would be worth the investment. The workshop decided to concentrate its plan on two systems—the proposed *WINDSAT* instrument and the next generation of satellite temperature sounders.

Several candidates were considered for use as the reference atmosphere—a time series of real atmosphere analyses or a long run of a general-circulation model. The latter was selected. The ECMWF agreed to generate the RA. It would be a 20-day prediction from a 15-level N48 resolution version of their grid-point model. In addition to the existing physics contained in the model, it would also include a diurnal cycle and the water-dimer effect. Once available, each of the participating agencies would conduct a series of fraternal-twin experiments with the RA using their own A/F models.

NMC's role in the effort was to generate a data base of simulated observations from the RA. Since the 20-day forecast period was chosen to coincide with the FGGE period of 10–30 November 1979, NMC simulated all available FGGE data during the period. In addition, they simulated observations from the proposed *WINDSAT* instrument. The complete data set was finished by NMC in late 1983. GLAS agreed to convert the NMC simulated data into FGGE format for use by each of the three centers.

In early 1984, GLAS had delivered the simulated data in FGGE format to NMC and had started their own experiments. NMC meanwhile reestablished a new long-range plan for conducting OSSE with a clear-cut goal of evaluating the potential contribution of *WINDSAT*. Although WPL had by now canceled its *WINDSAT* program, the Air Force interest was still alive, as was NASA's.

[4] Presently Goddard Laboratory for Atmospheres (GLA).

At the Remote Sensing Conference held during June 1984 in Clearwater Beach, Florida, several OSSE papers were presented by the NASA researchers. The first (Atlas et al., 1984) describes the current and planned OSSE activities of GLAS as part of the NMC-ECMWF-GLAS cooperative effort. The specific purpose was to design a simulation system which could be used to study the potential impact of advanced passive sounders and lidar temperature, pressure, humidity, and wind observing systems. The paper is notable in that it is the first to mention calibration. The authors point to a need to calibrate their results by comparison with real-data experiments performed with a similar system. Another consideration included in this paper that was not taken into account in previous work based on passive temperature profilers, is the effect of cloud cover. Here, an approach is developed, based upon the RA's relative-humidity values at various levels, which can be used with both passive sounders and active lidar systems. Finally, this paper makes a strong case for using a fraternal-twin experiment. Otherwise, as the authors point out, "the experiment may overestimate the skill of the forecast and underestimate the influence of the data on the analysis."

Halem and Dlouhy (1984) present the most recent results of the GLAS team in the first of two companion papers to Atlas et al. (1984). This paper concentrates on comparing a time series of 12-hourly forecasts using the GLAS A/F model with each of three RAs. The GLAS model itself is used to generate the first RA, permitting identical-twin experiments; the ECMWF model is used to generate the second RA, permitting fraternal-twin experiments; and a continuous sequence of NMC analyses is used as a third RA to allow for analysis-method experiments. The 12-hourly assimilation/forecast runs based on either an assimilation of perfect winds (complete, instantaneous, global, no observational error), perfect temperatures, and perfect surface pressure respectively are compared with each of the above RAs. Additional experiments consisting of the composite systems of wind and surface pressure, and temperature and surface pressure, were also compared with the RAs.

Among the general conclusions of this paper are 1) that the use of perfect winds alone has greater impact on model performance than perfect temperature data alone, 2) that all three RAs give similar results, suggesting that simulation-experiment results may not be model dependent as previously thought to be true, 3) that the model atmosphere adjusts to the wind in the extratropics as well as the tropics, and 4) that the adjustment times in the model are more rapid for wind than temperature. These experiments, however, were highly idealized, just as the title of the paper suggests, which clouds the interpretation of their results. Their conclusion regarding model adjustment to the insertion of tropical (30°N–30°S) winds, for example, seems to contradict both geostrophic adjustment theory (Monin and Obukhov, 1959; Washington, 1964; McPherson, 1975), which predicts that it is the motion field which will adjust to the mass field in the extratropics, and the results of such previous work as Gordon et al. (1972), and Williamson and Kasahara (1971).

The second paper (Dlouhy and Halem, 1984) presents a more-realistic simulation by considering global winds derived from a space-based lidar subject to three sources of error: 1) variations in aerosol content, 2) the presence of

clouds within the instruments' field of view, and 3) limited power for pumping the laser.

Using the GLAS RA discussed in their companion paper, Dlouhy and Halem considered the effects of reduced power and the presence of clouds by comparing simulated runs of each with a perfect wind simulation. Next, using the ECMWF RA they considered the effects of reduced power and limited aerosols by comparing simulated runs of each with a perfect wind simulation. Because these experiments combined identical- and fraternal-twin types, it is difficult to make adequate intercomparisons. The experiments are also still very contrived in that they do not include simulations of any conventional data. They did, however, represent the first attempts to simulate the effects of both clouds and aerosols on lidar winds.

4. Future experimental design considerations

The papers included in this review suggest that simulation can provide a reasonable means to evaluate the performance of proposed observing systems. We believe the future design of OSSEs should build upon the experience of others and the consensus which has formed from that experience. In this regard we need to consider the two major elements of the experiment, procedure and evaluation. The experimental procedure should include consideration of costs, observational-error structure, the experimental method, and calibration. The evaluation phase needs to consider which statistical parameters to use, statistical significance and the related question of data dependency, and consideration of the most appropriate data stratifications.

Cost cannot be overlooked in planning a simulation experiment if the purpose of the experiment is to determine a cost/benefit ratio between proposed observational-system costs and the potential benefits in savings expected from improved forecasts. If the estimated cost of the experiment exceeds the cost of the system, it would be wiser to simply fly the instrument and conduct actual data-impact studies. Cost of conducting the experiment will result from computer time used and manpower expended. Given the experiment design, these costs can be relatively easy to calculate and in turn used to trade off against modification in the experiment. These trades can be made as long as they do not result in critical components of either the procedure or the evaluation being compromised, such as inadequate number of computer runs, calibration, or statistical testing.

Virtually all researchers in this field agree that simulated observations must be given proper observational errors to make them appear as realistic as possible to an assimilation model. The use of perfect observations or poorly simulated errors can only lead to confusion in the interpretation of results. It is our belief that simulated observations should be obtained at the observation site and include as realistic an error structure as possible. Both random and systematic errors, including, where appropriate, the instruments' response to scales of motion not resolvable by our numerical A/F systems, should be included.

The choice of experimental method is also critical from the standpoint of proper interpretation of results. The consensus that we found is that a fraternal-twin or an analysis-method experiment should be conducted in preference to an identical twin. An experiment which includes a combination of methods only adds to confusion when attempting to make comparisons. Several papers we reviewed presented their results from experiments performed with an identical-twin procedure together with experiments based on a fraternal-twin procedure. It was obviously impossible to compare the performance changes in the one set with the other. Every attempt should be made to normalize the results so that once again confusion can be reduced. The RA and the assimilation model should stand in the same relationship to each other as the atmosphere and the model do. A model is always less sophisticated than the real atmosphere, so the assimilation model should therefore be less sophisticated than the RA. This can be accomplished by using a coarser grid and less physics.

Finally, some attempt should be made to calibrate the experiment so that the observed performance changes accompanying the inclusion or exclusion of a simulated data set in an assimilation and forecast can be estimated for the real data set. There is a potential difficulty which may arise when attempting to interpret the calibration of a proposed observation system which differs greatly, in its response to different scale of motion, from that of the calibration system. If rawinsonde-measured temperatures were used, for example, to calibrate a proposed satellite temperature sounder, it would be necessary to keep in mind that the two systems respond differently to different scales of motion. A simple calibration scheme, such as assuming a linear relationship between model-performance change in the real and simulated worlds, may be inadequate for such systems.

During the evaluation phase of an OSSE, the experimenter should select statistical parameters that have been widely accepted or, if new, should be ones that are readily interpreted. Since OSSEs based on large general-circulation models have been seen to rely most often on root-mean-square-error comparisons, their use in future experiments will have the added advantage that comparisons between studies can more readily be made. Other statistical measurements include the anomaly correlation, a statistic used to measure forecast skill relative to climatology. In addition to these old standards, new ways to measure forecast improvement, based on more economically important considerations such as storm-track forecasting improvements or computer flight-plan improvements, should be considered.

Although the purpose of the experiment may largely dictate which data stratifications will be used, there are certain ones which should be considered. These include composite statistics by wave group, by latitudinal region (equatorial or tropical versus mid and high latitude), by land versus ocean regions, by hemisphere, by level (surface, 500 mb, etc.), and by initial analysis time (00Z versus 12Z).

We have noted that the statistical significance of most results has not been satisfactorily addressed. Whenever statistical procedures are employed, it is essential to discuss their significance. Are the differences between the rms errors obtained from two simulation runs significant, or might they be treated as if they belonged to the same population? Common statistical techniques exist to determine significance between small samples. Small-sampling theory suggests that no fewer

than four or five forecasts should be used for this purpose. However, the larger the sample the more reliable the tests. A more difficult but related question is that of data dependency. And the question really is, what should be the minimum separation between forecasts? Two days is probably the absolute minimum as suggested earlier. However, to avoid the risk of dependence the experimenter should seek to separate forecasts by as large a period as possible. Questions related to statistical significance, although difficult to answer, must be addressed, in order that the experiment be considered complete.

Although GARP inspired much of the recent history of simulation, a strong need for continued research in this area is believed to exist. A highly reliable methodology based upon the guidelines suggested above will provide a very important service to decision makers and to the eventual quality of meteorological products.

Acknowledgments. The authors wish to thank Mrs. Mary Chapman who typed and helped edit the manuscript, Dr. John Brown for his support of this project, and Drs. Joe Gerrity and Ron McPherson for their many helpful suggestions and comments.

References

Alaka, M. A., and F. Lewis, 1967: Numerical experiments lending to the design of Optimum Global Meteorological Networks. Tech. Memo WBTM TDL-7, Environmental Science Services Administration, Weather Bureau, Washington, DC, 14 pp.

Atlas, R., M. Ghil, and M. Halem, 1981: Notes and Correspondence: Reply. *Mon. Wea. Rev.*, **109,** 201–204.

——, ——, and ——, 1982: The effect of model resolution and satellite sounding data on GLAS model forecasts. *Mon. Wea. Rev.*, **110,** 662–682.

——, E. Kalnay, J. Susskind, D. Reuter, W. E. Baker, and M. Halem, 1984: Simulation studies of the impact of advanced observing systems on numerical weather prediction. *Preprints Conference on Satellite Meteorology/Remote Sensing and Applications,* Clearwater, Fla., American Meteorological Society, 283–287.

Barwell, B. R., and A. C. Lorenc, 1985: A study of the impact of aircraft wind observations on a large-scale analysis and numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.*, **111,** 103–129.

Bentsson, L., and Morel, P., 1974: *Report on the Performance of Space Observing Systems for FGGE* GARP Working Group on Numerical Experimentation Report No. 6, WMO/ICSU, 12 pp.

Best, W. H., 1955: Differences in numerical prognoses resulting from differences in analyses. Tellus, **8,** 351–356.

Bristor, C. L., 1958: Effect of data coverage on the accuracy of 500 mb forecasts. *Mon. Wea. Rev.*, **86,** 299–308.

Bruce, R. E., L. D. Duncan, and J. H. Pierluissi, 1977: Experimental study of the relationship between radiosonde temperatures and satellite derived temperatures. *Mon. Wea. Rev.*, **105,** 493–496.

Cane, M. A., V. J. Cardone, M. Halem, and I. Halberstam, 1981: On the sensitivity of numerical weather prediction to remotely sensed marine surface wind data: a simulation study. *J. Geophys. Res.*, **86,** 8093–8106.

Charney, J., M. Halem, and R. Jastrow, 1969: Use of incomplete historical data to infer the present state of the atmosphere. J. Atmos. Sci., **26,** 1160–1163.

Dlouhy, R. and M. Halem, 1984: Observing system simulation experiments related to space-borne lidar wind profiling, Part 2: Sensitivity to atmospheric and instrumental influences. *Preprints, Conference on Satellite Meteorology/Remote Sensing and Applications,* Clearwater, Fla., American Meteorological Society, 280–282.

Gandin, L. S., M. A. Alaka, S. A. Mashkovitch, and F. Lewis, 1967: *Design of Optimum Networks for Aerological Observing Stations.* World Weather Watch Planning Report No. 21, World Meteorological Organization, Geneva, Switzerland, 58 pp.

Gordon, C. T., L. Unscheid, and K. Miyakoda, 1972: Simulation experiments for determining wind data requirements in the tropics. *J. Atmos. Sci.*, **29,** 1064–1075.

Halem, M. and R. Jastrow, 1970: Analysis of GARP data requirements. *J. Atmos. Sci.*, **27,** 177.

——, and R. Dlouhy, 1984: Observing system simulation experiments related to space-borne lidar wind profiling. Part 1: Forecast impacts of highly idealized observing systems. *Preprints, Conference on Satellite Meteorology/Remote Sensing and Applications,* Clearwater, Fla., American Meteorological Society, 272–279.

Huffaker, R. M., 1978: Feasibility study of satellite-borne lidar global wind monitoring system. NOAA Tech. Memo. ERL WPL-37, 297 pp.

Jastrow, R. and M. Halem, 1970: Simulation studies related to GARP. *Bull. Amer. Meteor. Soc.*, **51,** 490–513.

——, and ——, 1973: Simulation studies and the design of the First GARP Global Experiment. *Bull. Amer. Meteor. Soc.*, **54,** 13–21.

Jess, E. O., 1959: A numerical prediction experiment involving data faucity and random errors. *Tellus,* **12,** 21–30.

Kasahara, A., 1972: Simulation experiments for meteorological observing systems for GARP. *Bull. Amer. Meteor. Soc.*, **53,** 252–264.

——, and D. Williamson, 1972: Evaluation of tropical wind and reference pressure measurements: numerical experiments for observing systems. *Tellus,* **24,** 100–115.

Koehler, T. L., J. C. Derber, B. D. Schmidt, and L. H. Horn, 1983: An evaluation of soundings, analyses and model forecasts derived from TIROS-N and NOAA-6 satellite data. *Mon. Wea. Rev.*, **111,** 562–571.

Lorenc, A. C., 1975: Results of observing systems simulation experiments for the First GARP Global Experiment. *GARP Working Group on Numerical Experimentation, Report #10,* 37–68.

McPherson, R., 1975: Progress, problems, and prospects in meteorological data assimilation. *Bull. Amer. Meteor. Soc.*, **56,** 1154–1166.

Monin, A., and A. Obukhov, 1959: A note on the general classification of motions in a buroclinic atmosphere. *Tellus,* **11,** 159–162.

Morel, P., G. Leteure, and G. Rabreaa, 1971: On initialization and non-synoptic data assimilation. *Tellus,* **23,** 197–206.

Newton, C. W., 1954: Analysis and Data Problems in Relation to Numerical Prediction. *Bull. Amer. Meteor. Soc.*, **35,** 287–294.

Nitta, T., 1975: Some analyses of observing systems simulation experiments in relation to the First GARP Global Experiment. *GARP Working Group on Numerical Experimentation, Report #10,* 1–35. Plan for U.S. Participation in the Global Atmospheric Research Program, National Academy of Sciences, Washington, DC, 1969.

Ohring, G., 1979: Impact of satellite temperature sounding data on weather forecasts. *Bull. Amer. Meteor. Soc.*, **59,** 1225–1240.

Schlatter, T. W., 1981: An assessment of operational TIROS-N temperature retrievals over the United States. *Mon. Wea. Rev.*, **109,** 110–119.

Smagorinsky, J., K. Miyakoda, and R. F. Strickler, 1970: The relative importance of variables in initial conditions for dynamical weather prediction. *Tellus,* **22,** 141–157.

Tracton, M. S., A. J. Desmarais, R. J. Van Haaren, and R. D. McPherson, 1980: The impact of satellite soundings on the National Meteorological Center's analysis and forecast system—the data systems test results. *Mon. Wea. Rev.*, **108,** 543–586.

——, ——, ——, 1981: On the system dependency of satellite sounding impact—comments on recent impact test results. *Mon. Wea. Rev.*, **109**, 197–200.

U.S. Committee for GARP, 1969a: *Plan for United States Participation in the Global Atmospheric Research Program.* National Academy of Sciences, Washington, D.C., 25 pp.

Warbuton, J. D., 1984: *Visit to European Center for Medium-Range Weather Forecasting.* EOARD Report, EOARD LR-894-07, European Office of Aerospace Research and Development, Air Force Systems Command, Bolling AFB, Washington, DC, 2 pp.

Washington, W., 1964: A note on the adjustment towards geostrophic equilibrium in a simple fluid system. *Tellus*, **16**, 530–534.

Wilcox, R. W., and F. Sanders, 1976: Comparison of layer thickness as observed by nimbus E microwave spectrometer and by radiosonde. *J. Appl. Meteorol.*, **15**, 956–961.

Williamson, D. L., 1973: The effects of forecast error accumulation on four-dimensional data assimilation. *J. Atmos. Sci.*, **30**, 537–543.

——, and A. Kasahara, 1971: Adaption of meteorological variables forced by updating. *J. Atmos. Sci.*, **28**, 1313–1324.

Yu, T. W., and R. D. McPherson, 1984: Global data assimilation experiments with scatterometer winds from SEASAT-A. *Mon. Wea. Rev.*, **112**, 368–376. •

announcements[1]

meetings of interest

25–29 August 1986. A workshop, titled "Scaling, Fractuals and Nonlinear Variability in Geophysics, Fundamentals and Applications," will be held from 25 to 29 August 1986 at McGill University, Montreal, Canada. The workshop will confront theories and experiments on the scaling behavior of geophysical fields. Particular emphasis will be given to turbulence in the atmosphere and oceans, and to the implications of scaling to models and in situ and remotely sensed measurements. For more information, contact either D. Schertzer or S. Lovejoy, Physics Department, McGill University, 3600 University St., Montreal, Quebec H3A 2Y8, Canada; telephone (514) 392-5135, 4405.

8–10 September 1986. An international specialty conference titled, "Visibility Protection: Research and Policy Aspects," will be held from 8 to 10 September 1986 at the Jackson Lake Lodge in Grand Teton National Park, Wyoming. The conference will focus on major research topics as well as regulatory and policy aspects of visibility protection. Topics to be discussed include regulatory and policy formulation; measurements; modeling; source attribution; urban visibility; atmospheric optics; aerosols; perception benefit analysis; and emission inventories. A special section devoted to the linkage between visibility protection and mitigation of acid deposition will also be included. For further information contact the Air Pollution Control Association, Meetings Department, P.O. Box 2816, Pittsburgh, PA 15230; telephone (412) 232-3444.

6–10 April 1987. The 16th international technical meeting on air pollution modeling and its application will be held in Lindau, Federal Republic of Germany, from 6 to 10 April 1987. The meeting will focus on models for the large-scale transport and deposition processes of air pollution including acidification and heavy metals. Key topics include dry deposition; theory and experiment; wet scavenging processes and physico/chemical processes in clouds; meteorological parameterization in dispersion modeling; model verification and policy implications; and new developments in dispersion modeling and theory. For further information, contact Han van Dop, KNMI, P.O. Box 21, 3730 AE DE BILT, The Netherlands.

24 August–4 September 1987. The 22nd General Assembly of the International Union of Radio Science will be held in Tel Aviv, Israel, from 24 August to 4 September 1987. For more information, contact J. Shapira, Organizing Committee, Secretariat, P.O. Box 50006, Tel Aviv, Israel.

9–22 August 1987. A symposium, titled "Forest Hydrology and Watershed Management," will be held as part of the International Association of Geodesy and Geophysics XIX General Assembly from 9 to 22 August 1986 in Vancouver, British Columbia, Canada. Principal theme topics include the presentation of case studies of applied watershed management; discussions relating use of, or attempts to use, hydrologic models to extrapolate research results; presentations on the effects of forests on the chemistry and quality of runoff from catchments receiving acid precipitation; and discussions of hydrological processes either influenced by or that have an influence on woody vegetation. For more information contact R. H. Swanson, Principal Convener, IAHS Symposium on Watershed Management, Northern Forestry Centre, 5320-122nd Street, Edmonton, Alberta, Canada T6H3S5.

14–15 August 1987. A half-day workshop and a full-day field trip on debris torrents will be held by the International Association of Hydrological Sciences program as part of the International Union of Geodesy and Geophysics XIX General Assembly in Vancouver, British Columbia. The objective of the workshop will be to review theoretical developments in the understanding of debris torrents (debris flows) as a hydrologic phenomenon. Oral or poster papers should be sent by 30 November 1986 to Olav Slaymaker, Department of Geography, University of British Columbia, Vancouver, B.C. V6T 1W5. For more general information contact G. J. Young, CNC/IAHS, Inland Waters Directorate, Environment Canada, Ottawa, Ontario K1A OE7.

[1] Notice of registration deadlines for meetings, workshops, and seminars, deadlines for submittal of abstracts or papers to be presented at meetings, and deadlines for grants, proposals, awards, nominations, and fellowships must be received at least three months before deadline dates.—*News Ed.*

(*continued on page 712*)