Summary of Breakout Session on Metrics and Validation

Metrics and validation represent two different activities of the Community Coordinated Modeling Center:

- The term "model validation" refers to a broad effort to test and exercise models, for a wide range of circumstances and applications. CCMC efforts at validation involve a variety of tests and studies carried out by CCMC personnel, who acquire considerable hands-on experience with the codes. To exercise and test the models more widely, CCMC also makes run results available to the scientific community and does model runs on request. The totality of information accumulated in these validation studies is intended to help judge when a model is ready for transition to a rapid prototyping center.
- A "metric" is a single number that is used to indicate the agreement between different models or algorithm and observations. It can also be used as a quantitative index of the progress of a field of research. A metric provides an objective though narrow measure of model performance, and it can be applied to a large number of models. The initial organized effort at space weather metrics, the GEM Metrics Challenge, was a competition involving a substantial number of magnetospheric models and algorithms. The models were run by their developers, but CCMC acted as impartial judge.

The discussions of the session focused much more on metrics than on validation, partly because the use of metrics is particularly controversial in the space weather science community. The controversy does not center on the CCMC but on broader questions about the scientific usefulness of metrics and how to make future competitions as fair and useful as possible. There is no corresponding controversy about the need for model validation, and validation is unquestionably a central activity of CCMC. Thus the discussion in the breakout session emphasized questions about metrics more than validation. It should be noted, however, that some issues are common to metrics and validation. An important example is the problem of quality control of input data.

The breakout session included a number of invited presentations. Bob Robinson introduced the idea of metrics and explained that the National Space Weather Program needs a few simple metrics to measure progress of space weather science. Dick Wolf and Bob Schunk talked about magnetosphere-ionosphere and ionosphere-thermosphere metrics, respectively, as scientists who participated in the Study of Space Weather Metrics. NSF recently asked the SHINE Steering Committee to suggest a solar-interplanetary metric, and Janet Luhmann reported on the results of those deliberations. Steve Quigley described the Air Force rapid prototyping effort and discussed DoD concerns about metrics and validation. Terry Onsager gave a corresponding report for NOAA. Michael Hesse spoke on the role of CCMC role in metrics and validation and on the experience with the GEM Metric Challenge.

Jan Sojka led discussions, which turned out to be vigorous and sometimes heated. However, consensus seemed to be achieved on some issues:

- The recent GEM Metrics Competition was a valuable first effort at evaluating space weather metrics. Practical lessons were learned on how to make future contests more efficient and meaningful.
- Future metrics competitions must be "blind," with no opportunity for modelers to adjust their models to fit the data. Because the initial GEM competition was not blind, results from it cannot be used to judge the relative accuracy of the different models.
- CCMC should continue to serve as unbiased judge of metrics competitions. Since CCMC does not have the resources to run all of the competing models, a way must be found for blind model runs to be carried out at modelers' institutions.
- The present first-priority NSWP thermosphere-ionosphere-magnetosphere metrics (based on high-latitude ionospheric electric fields and global electron densities) represent reasonable first efforts at space weather metrics and should be minimally sufficient to satisfy the administrative need to quantify progress of the NSWP over the remainder of its lifetime. However, no single metric can be broad enough to provide an adequate characterization of a major area of space-weather science.
- Though they represent reasonable first efforts, the present NSWP magnetospheric and ionospheric metrics are not optimal and sometimes give results that conflict with scientific judgment.
- Whenever possible, metrics should be defined with direct input from the users of

space weather products.

- Metrics should be evaluated routinely and for longer time periods than were used for the initial GEM metrics competition.
- More effort is needed to assure the cleanliness and accuracy of the observational data for metrics competitions. The observers should be asked to specify error bars, if technically possible.

There were a number of significant points on which there was no clear consensus, including the following:

- It is not clear whether the state of solar-heliospheric modeling has reached a state of maturity where a quantitative metric is meaningful. The general difficulty of characterizing the state of a whole area of science by one number is particularly evident in the solar-heliospheric area.
- Because it was not a blind test, it is not clear whether results of the GEM Metrics Challenge represent a scientifically valid first point on the long-time curve that will document the progress of NSWP.
- The value of keeping scientific metrics distinct from application metrics is not clear. For example, the metric used for the Electrojet Challenge, which was aimed directly at a user needs, may be as scientifically meaningful as the presently adopted NSWP ionosphere-magnetosphere metric, which was based on high-latitude ionospheric electric fields. Perhaps the electrojet-challenge metric should be implemented in future CCMC-judged magnetosphere-ionosphere metric challenges.

- Though the present NSWP ionospheric and magnetospheric metrics could be improved for better conformity with scientific judgment, it is not clear whether or not that would be worth the effort.
- It would be possible to develop more sophisticated multi-level sets of metrics that would be more diagnostic of the science. It is not clear whether or not that would be worth the effort.
- There are several promising approaches to allowing metric challenges to operate efficiently over longer time periods, but it is not clear which is best. One possibility involves running a test one day per month (e.g. world day). Another involves running successive models through the same large "clean" dataset supplied by an operating agency, for which the "right answers" are kept confidential. There are possibilities for automating competitions.
- There was no consensus on who should be responsible for cleaning input data sets (i.e., identifying and removing bad data points) for future metric competitions or for validation studies. Should each model clean each of its input data sets, or should the cleaning algorithms be supplied by experts on the instruments involved? Or a combination of both?

Models play an increasingly important role in both the research and operational aspects of space weather. Space weather models must be carefully and objectively evaluated, both to measure the overall progress of the NSWP and to aid in decisions about which models should be implemented at operational centers. Metrics and validation are essential parts of the evaluation process but are still somewhat new to the space science

community. Some experimentation will be necessary to find the best approaches, but it is clear that CCMC has a unique and central role in the process.