Measuring the Performance of Scientific Models

K. A. Keller, M. Hesse, L. Rastaetter,
M. M. Kuznetsova, T. Moretto
NASA Goddard Space Flight Center

Abstract

The use of physics-based models to forecast space weather will require that these models be validated. One tool for measuring the progress of the space weather models is metrics. A scientific metric as defined by the United States National Space Weather Program has three elements: 1). An output parameter from the model such as currents in the ionosphere, 2.) A satellite or ground-based measurement that can be used for comparison, and 3.) A quantifiable parameter that can measure the difference between the model parameter and the measurement.

The Community Coordinated Modeling Center (CCMC) has used ionospheric currents from empirical models to compute ground magnetic perturbations and compared the results to data obtained from ground magnetometer chains. We will present results from these metric studies. CCMC is also using geosynchronous data to validate inner magnetosphere models. We will discuss the initial results from this study and future plans for metric studies for these models. In addition, we discuss plans for validating global MHD models.

Need for Metrics

- Create objective measure of current capabilities both for scientific and operational needs.
- Measure the improvement of model capabilities over time.
- Provide an objective comparison between models with comparable output.

Metrics which lead to scores near unity now are useless!

Elements of a Metric

- An output parameter from a model.
  - An example is currents in the ionosphere can be used to calculate ground magnetic perturbations.
- A satellite or ground-based measurement that can be used for comparison.
  - An example is ground magnetometer data.
- A quantifiable norm that assesses the difference between the parameter from the model and the measurement.

Possible Metrics

- Ground magnetic perturbations using data from ground magnetometer chains.
- Particle fluxes at geosynchronous orbits using Los Alamos National Laboratory satellite data.
- Other metrics that may be suggested by the space weather operational or research community.

Community Coordinated Modeling Center (CCMC)

- Multi-agency partnership established to help bridge the gap between the space weather research community and operational agencies of National Oceanographic and Atmospheric Administration and the United States Department of Defense.
- Provides validation of models through both science-based testing and metrics evaluations by an independent evaluator.
- Serves the space weather research community by providing access to models through runs-on-request web site.

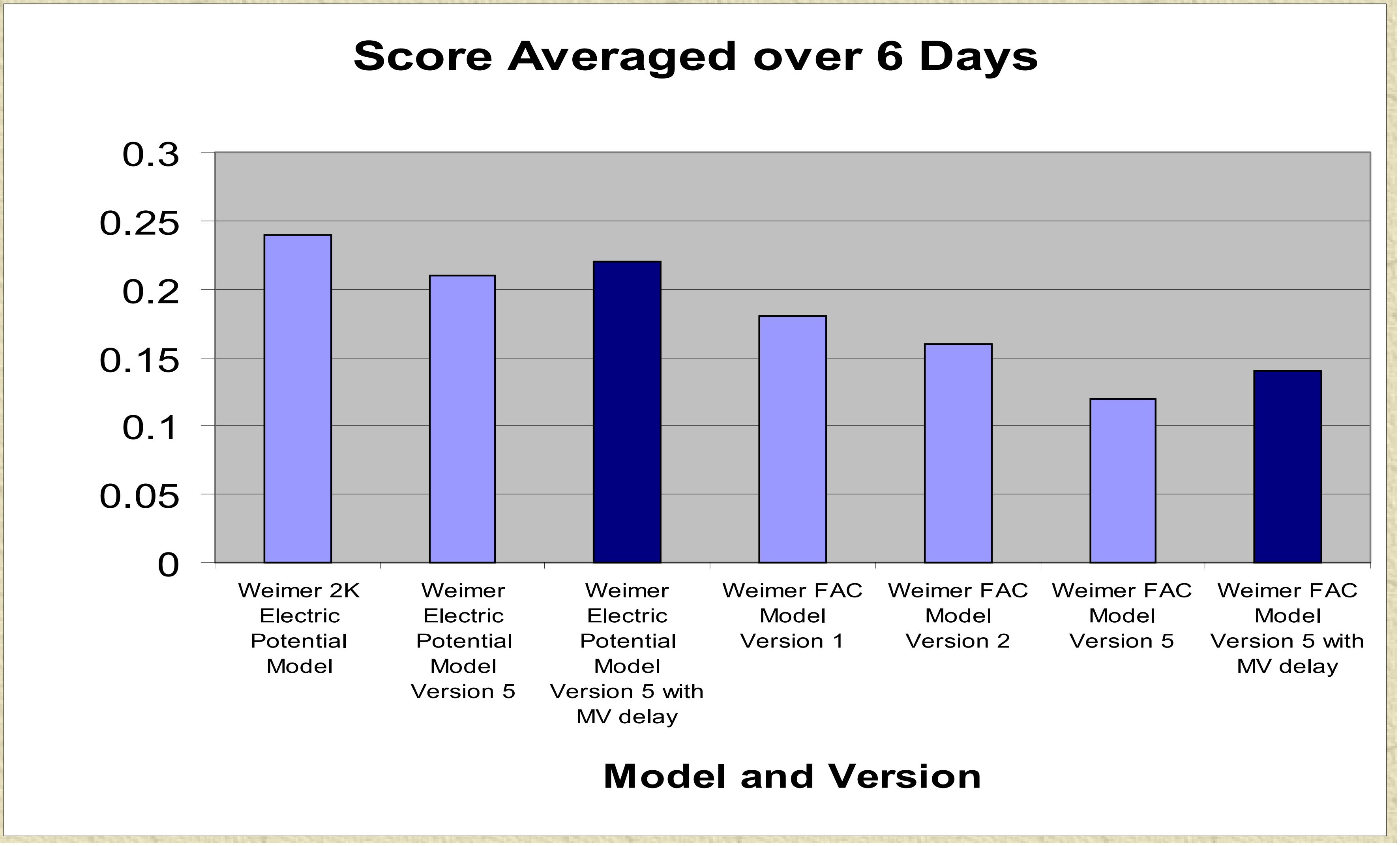
Ground Magnetic Perturbations

- Data
  - 10 stations in the Greenland chain using the H component of the data.
- Models
  - Weimer electric potential model (2 different versions).
  - Weimer field-aligned current model (3 different versions).
- Skill score
  - An individual model is scored $D_i = \Sigma|\Delta H_{model} - \Delta H_{data}|/npts$.
  - A skill score is computed for each ground station by

$$M_i = 1 - D_i / D_s$$

where D_s is for the standard model. In this case, the standard model is $\Delta H_{standard} \equiv 0$.

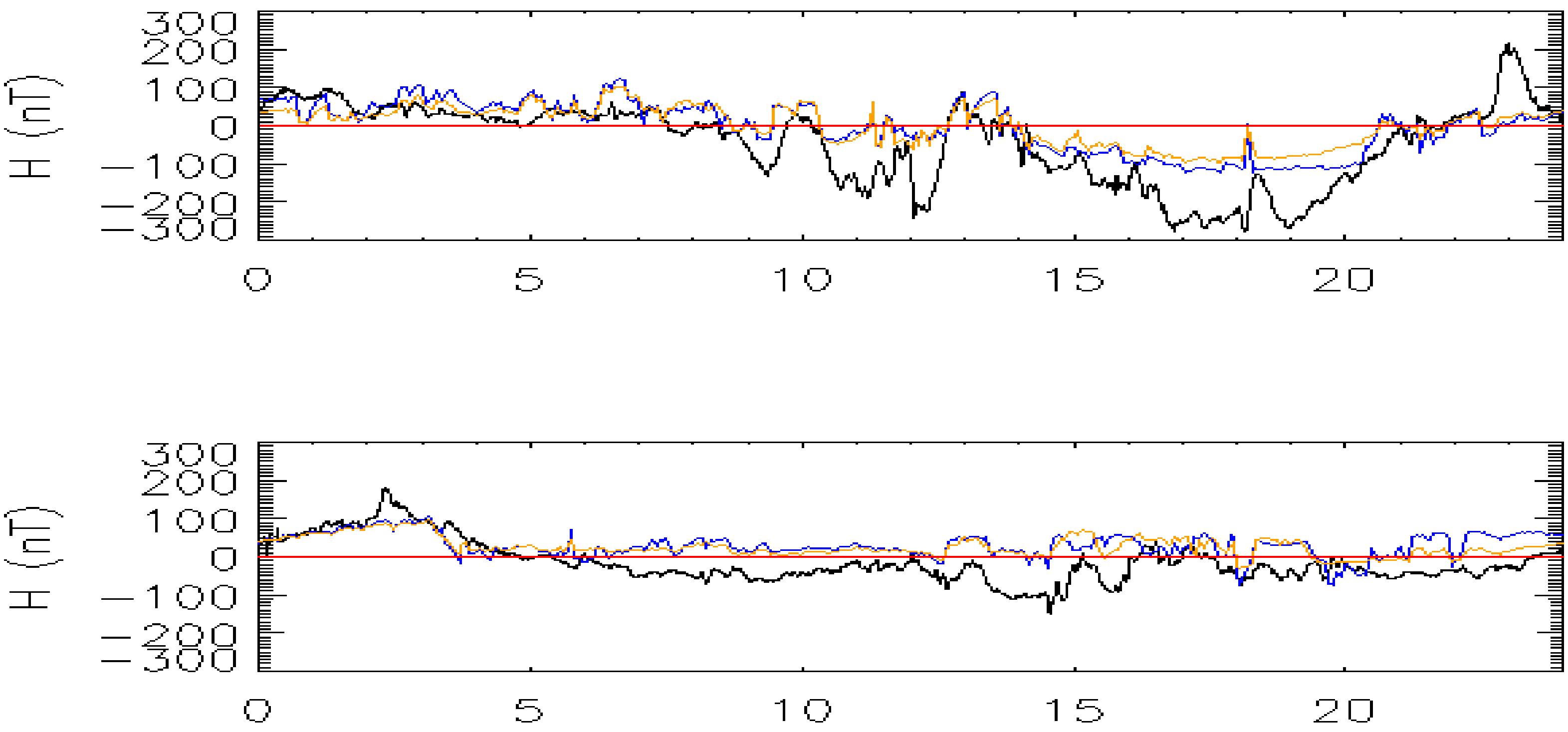
Results for Weimer Models (averaged over 10 stations) for H component.



Parameter Tests

- Different time delays for the ACE data were used. The skill scores were not very sensitive to the time delays. There was a slight improvement when using minimum variance technique received from Dan Weimer.
- Different Hall conductivities were used for the electric potential model. The skill scores were better for Hall conductivities of 5 and 7.5 mhos. For later versions, the scores are more sensitive to different conductivities.

Comparison of Model Results to Data



Black: Data from ground magnetometers

Orange: Model results from Weimer 2k Electric Potential Model

Blue: Model results from Weimer Electric Potential Model Version 5

Magnetometer data was provided by the Danish Meteorological Institute (Dr. Jurgen Watermann, Project Scientist)

Comparison of Model Results to Data Discussion

- In the top plot, the results from the Weimer 2K electric potential model tend to be smaller in magnitude than the results from Weimer electric potential model version 5. Since the results have the same sign as the data, the score for the version 5 model is better for this station on this day. Both scores are in the .2 -.3 range.
- In the bottom plot, the results from the 2K version again tend to be smaller in magnitude than the results from version 5 model. On this day, there is significant periods of time when the model has the wrong sign compared to the data. In this case, the score for the 2K version is better. The scores for this station and day are either negative or around zero.
- For each day, there is at least one station with the wrong sign for a significant period of time. Since the 2K version tends to predict smaller magnitudes, it tends to do better when the sign is incorrect. This tends to give better scores for the 2K version when the scores are averaged over 10 stations.

Future Plans

- Currently, we are using only the Greenland chain. This gives a range of stations in latitude but is limited in local time. We want to add stations that would give us a broader coverage of local times.
- We will do similar tests for MHD models.

Proton Fluxes

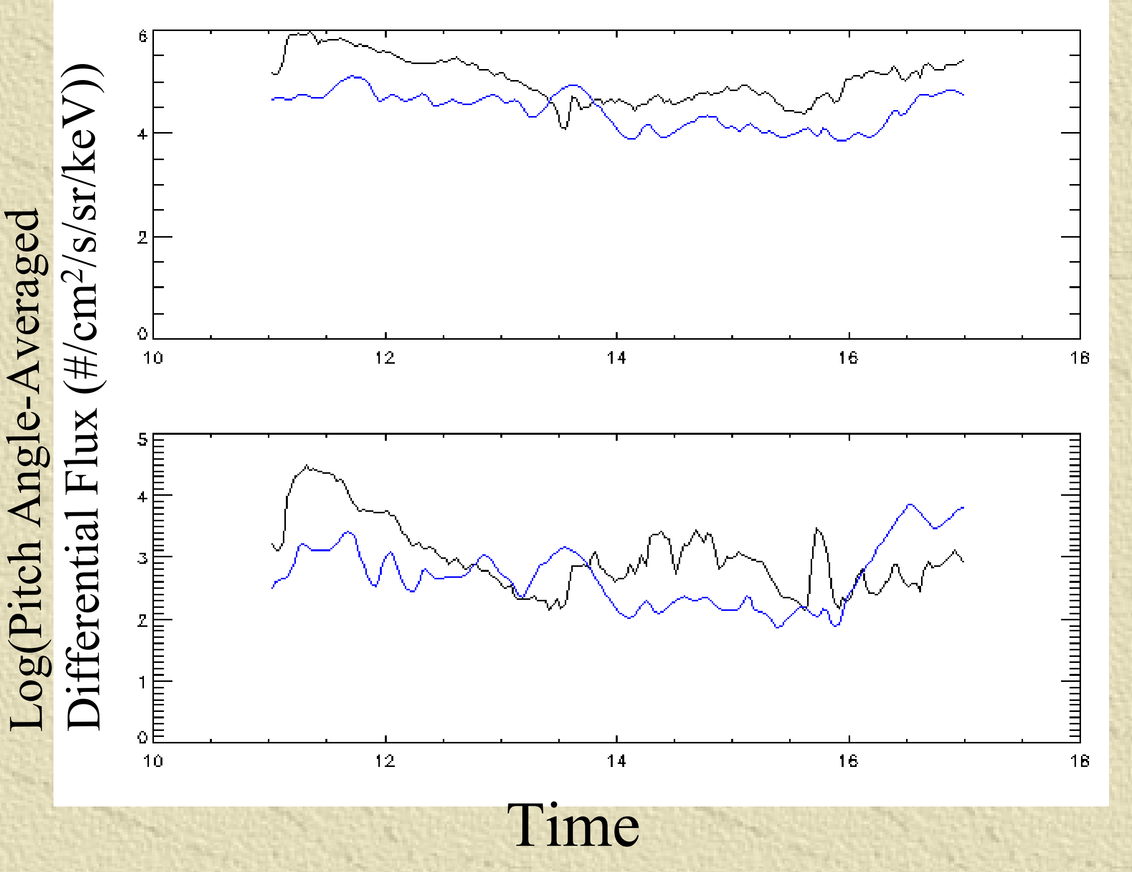
- Data
  - Proton fluxes from LANL geosynchronous satellites
- Model
  - Fok ring current model coupled to MHD models
- Skill Score using the mean square error
  - Calculate mean square error

    MSE = $(\sum(\text{predicted} - \text{observed})^2/\text{npts})$
  - Calculate variance of observations

    STD = $(\sum(\text{observed} - \text{mean})^2/\text{npts})$
  - Skill score

    Skill score = 1- MSE/STD
- Cross Correlation

Sample of Ring Current Metric



Energy Band (keV)	Skill Score	Cross Correlation
50-75	0.135	.59
113-170	-0.02	0.07

Black is LANL data. Blue is the model results.

Geosynchronous proton flux data was provided by the Energetic Particle team at Los Alamos National Laboratory, Richard Belian (PI).

Discussion

- The plot contains two samples at different energies (lower energy is on top) over a 6-hour period. Black is the data and blue is the model results. The graphs are logs of the pitch-averaged differential flux. The actual scores compare the fluxes and not the logs of the fluxes.
- Neither scored particularly high on the skill score. The lower energy bin scored higher using the cross correlation.
- Besides metrics, we have done scientific validation studies on the ring current model. The ring current model matches the data best when the increase or decrease in flux is directly driven by changes in the solar wind.

Future Plans for Inner Magnetosphere Models

- We plan to do the skill score using several different energy bands for different days and 2-3 satellites per day.
- We will do the same comparison using electron data at the same energies. In this case, we will test two different versions of the Fok ring current model. These models use different density and temperature profiles.
- We will also do comparisons for higher energies with the Fok radiation belt model.

Future Plans for Global MHD Models

- Metric using ground magnetometer data to test ionospheric currents
- Community wide metrics
  - To be determined by the community
  - Possible candidates
    - Comparison with DMSP satellites
    - Comparison with GOES data
- Metrics for inner magnetosphere models coupled to global MHD models

Summary

- The ground magnetic perturbations is a first attempt at creation and application of a standard and repeatable metric.
- Blind test (no fine tuning)!
- Fine tuning of metrics is required in collaboration with the operational agencies and researchers.
- First steps, more to come.