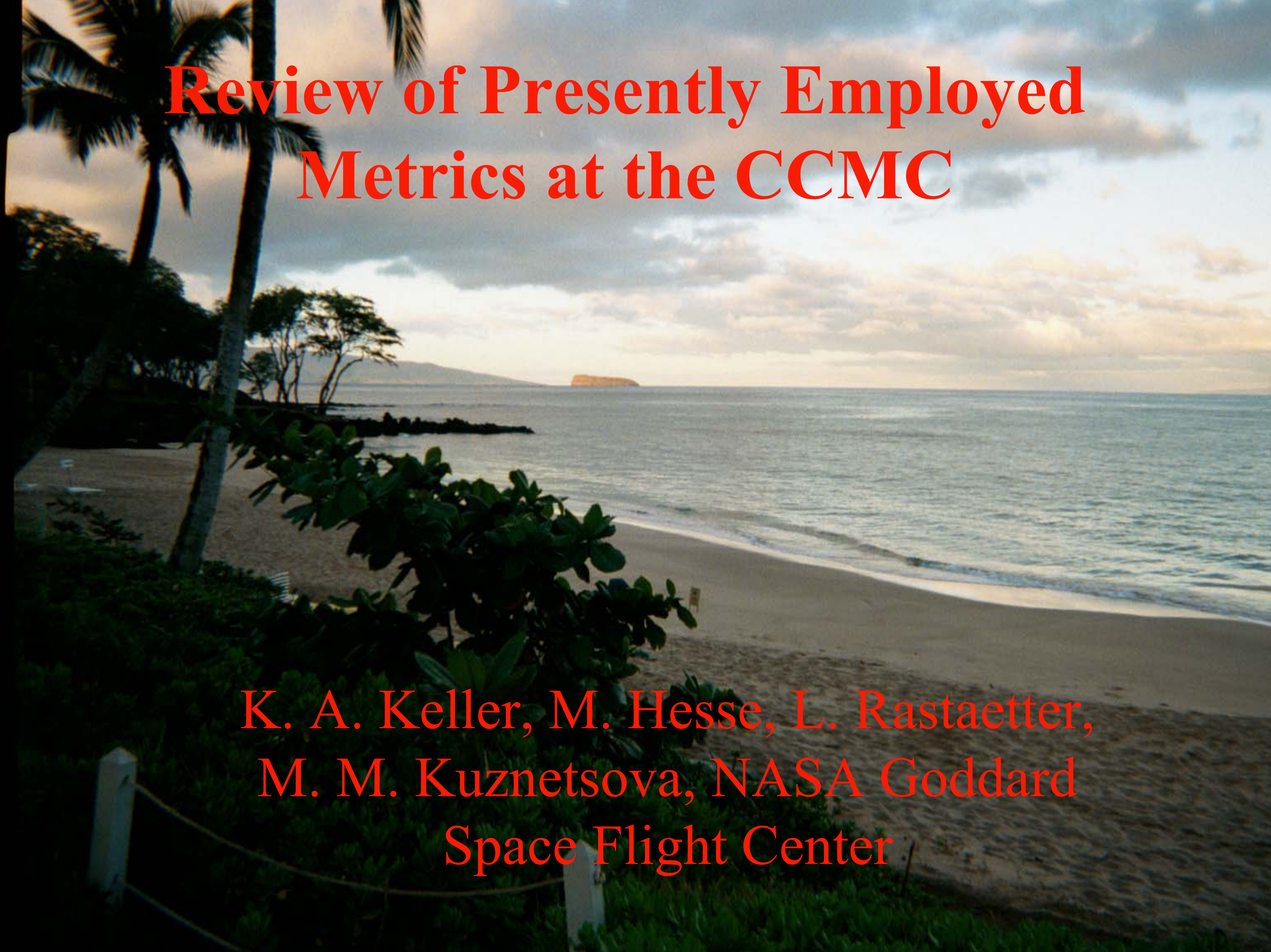


Review of Presently Employed Metrics at the CCMC

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

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.

Long-Term Metric Validation

- A long-term data source like ground-based magnetometers or LANL geosynchronous satellite data is needed.
- CCMC will keep a list of days/events and rerun these days/events with each new model or new version of model.
- CCMC will manage a database of scores and present these results to the community.

Other Model Validation

- CCMC also uses metrics to test the sensitivity of models to parameters.
  - For the Weimer ionosphere models, we have used metrics to test the sensitivity of different time delays for solar wind propagation and Hall conductivity.
  - For the Fok Ring Current Model, we will work with Mei-Ching Fok to optimize and validate the empirical formula in the radiation belt model using metrics.
- Community-wide metric evaluation for large events.
  - Community wide participation will enlarge the number of models being tested.
  - Multiple metrics can be tested by the community.

Current Metrics

- Ground magnetic perturbations using data from ground magnetometer chains.
- Particle fluxes at geosynchronous orbits using Los Alamos National Laboratory satellite data.

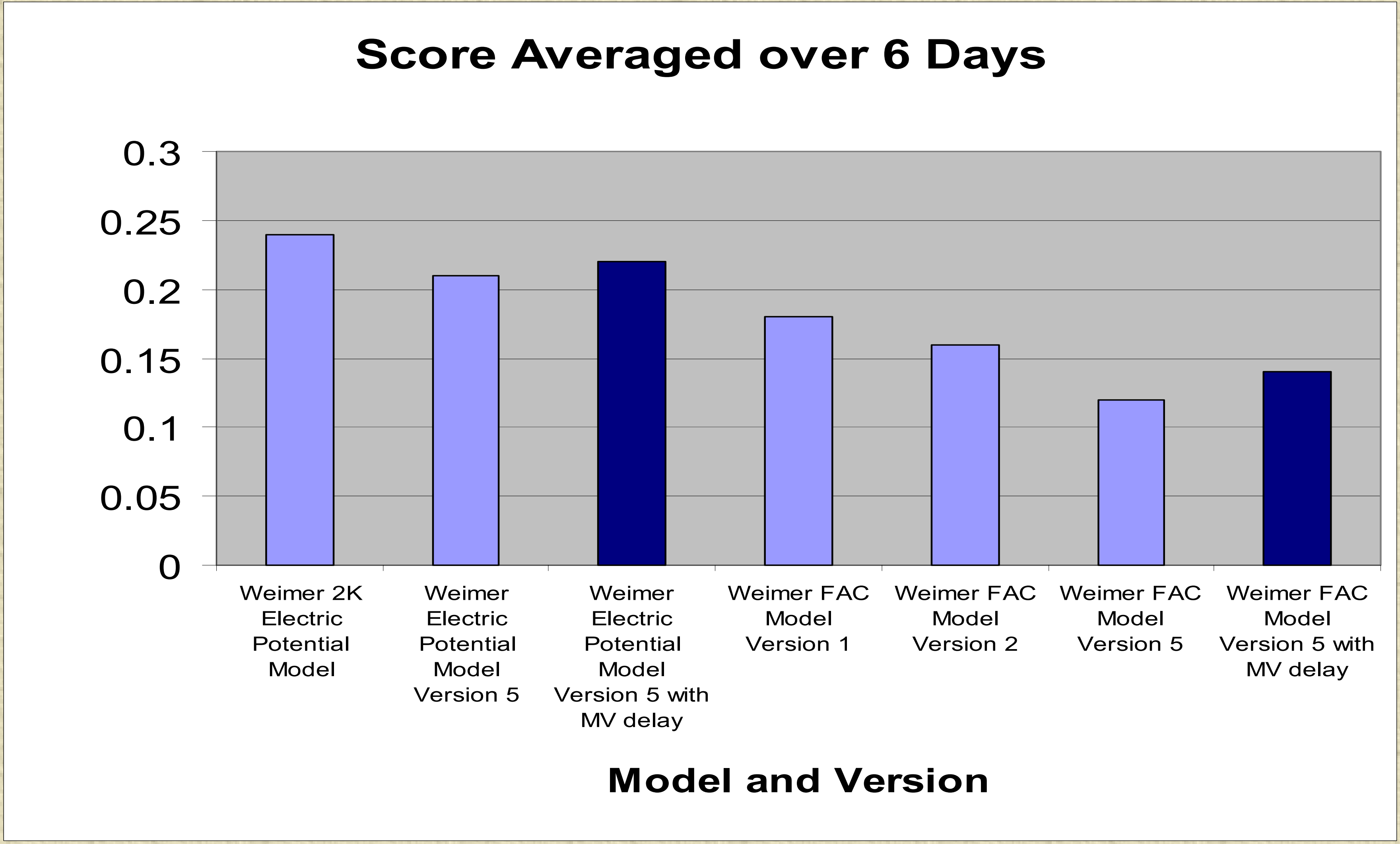
Ground Magnetic Perturbations

- 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 $\Delta H_{standard} \equiv 0$.

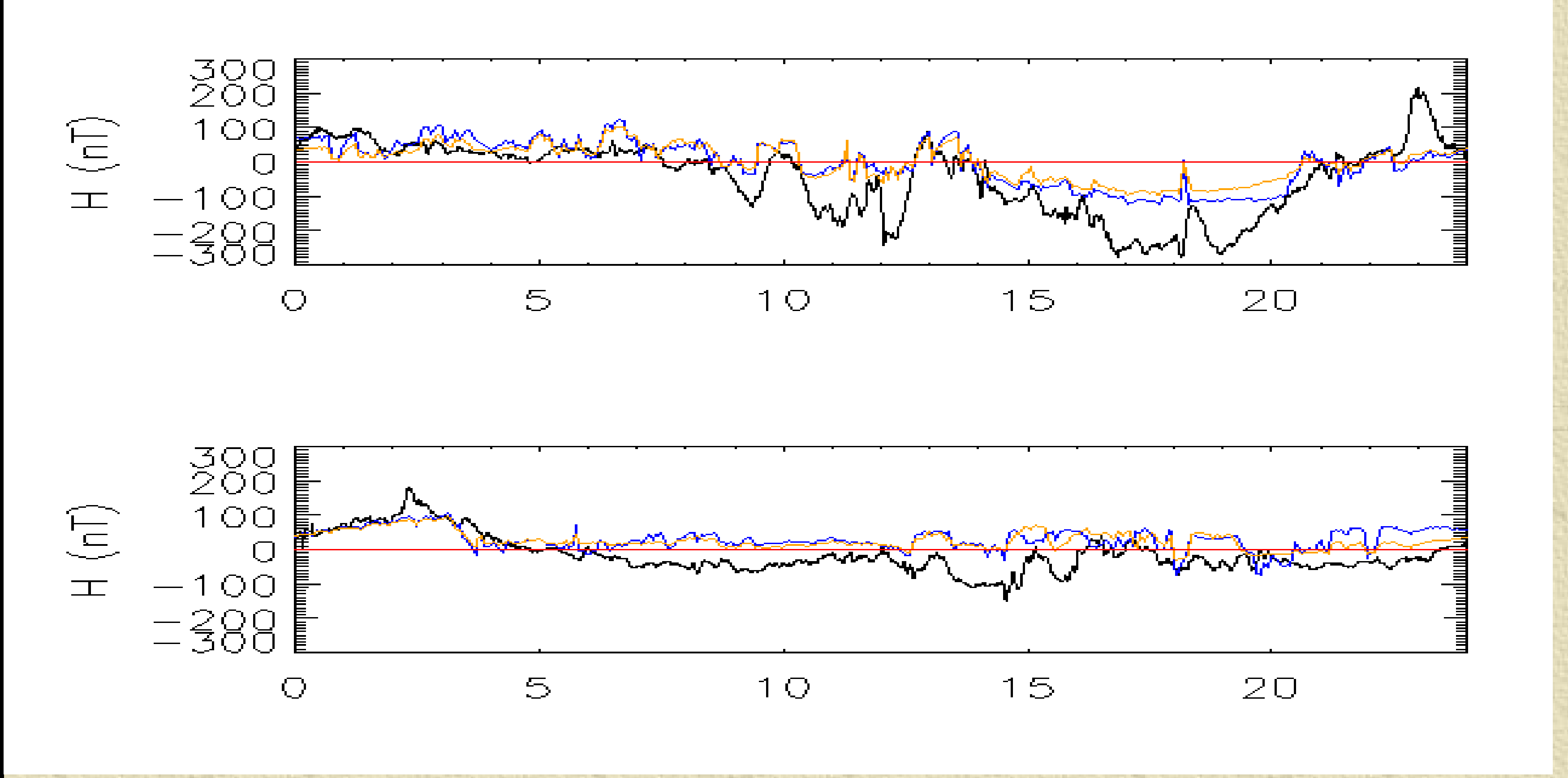
Results for Weimer Models (averaged over 10 stations) for H component.



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 driven by MHD models

Skill Scores Determination as Defined by SMC

- Skill Score using the Log Mean Square Error
  - Calculate log (mean square error)

    LogRMSE = sqrt{$\sum$(log $_{10}$ |predicted – observed|)^2/npts}
  - Calculate log (variance of observations)

    Log STD = sqrt{$\sum$(log $_{10}$ |observed – mean|)^2/npts}
  - Skill score

    Skill score = 1- logMSE/logSTD
- Skill Score using the Root Mean Square Deviation
  - Calculate mean square error

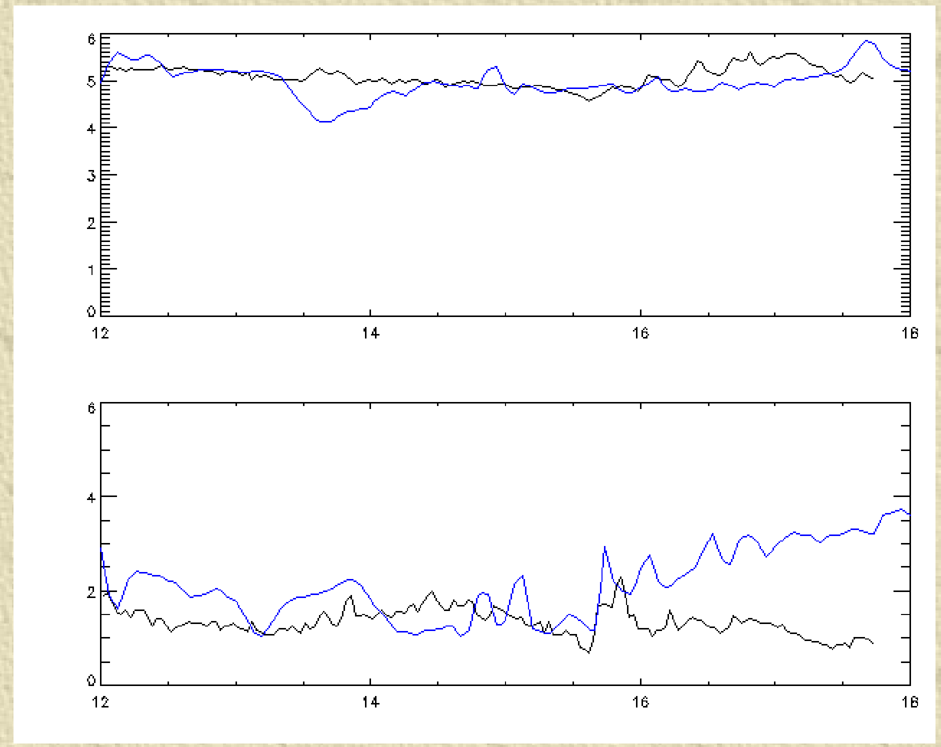
    RMS_deviation = {$\sum$(log $_{10}$ [predicted /observed])^2/npts}
  - Calculate variance of observations

    STD_deviation = {$\sum$(log $_{10}$ [observed /mean])^2/npts}
  - Skill score

    Skill score = 1- RMS_deviation/STD_deviation

Sample of Ring Current Skill Scores Storm Day

Log(Pitch Angle-Averaged Differential Flux (#/cm^2/s/sr/keV))



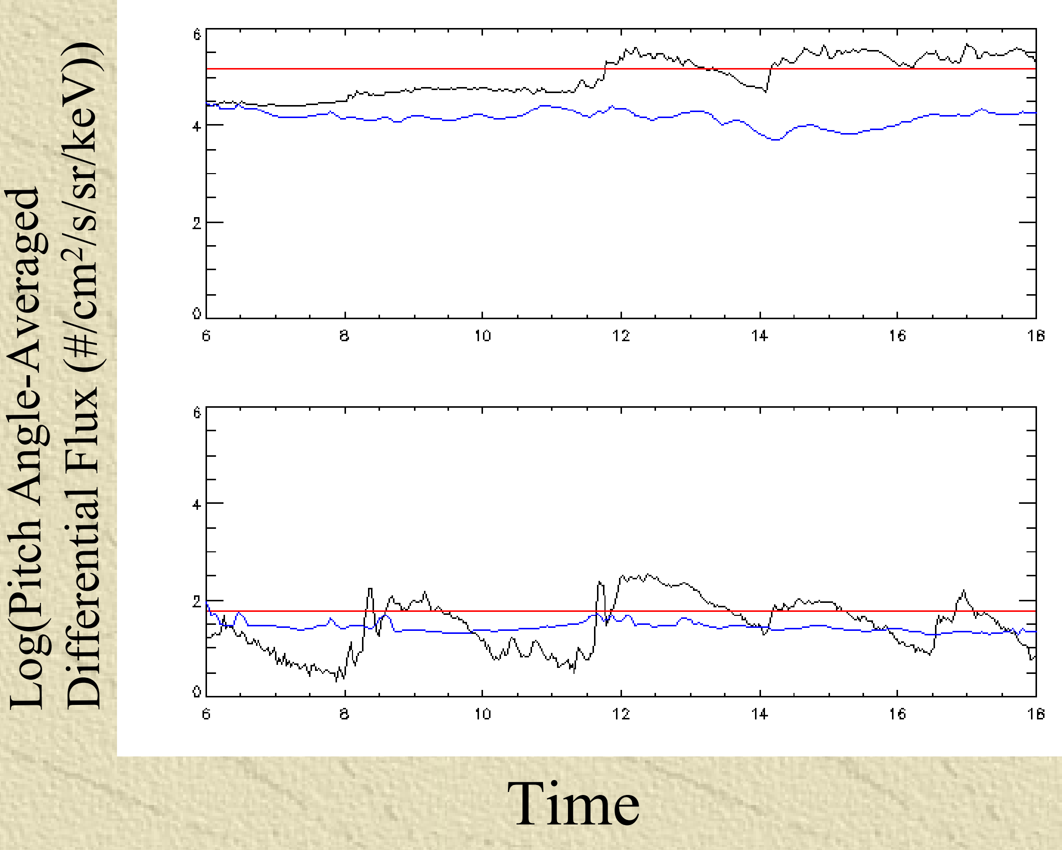
Time

Black is LANL data. Blue is the model results.

Energy Band (keV)	Log Mean Square	Root Mean Square
50-75	0.0038	.43
250-400	-0.946	-.49

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

Sample of Ring Current Skill Scores Sawtooth

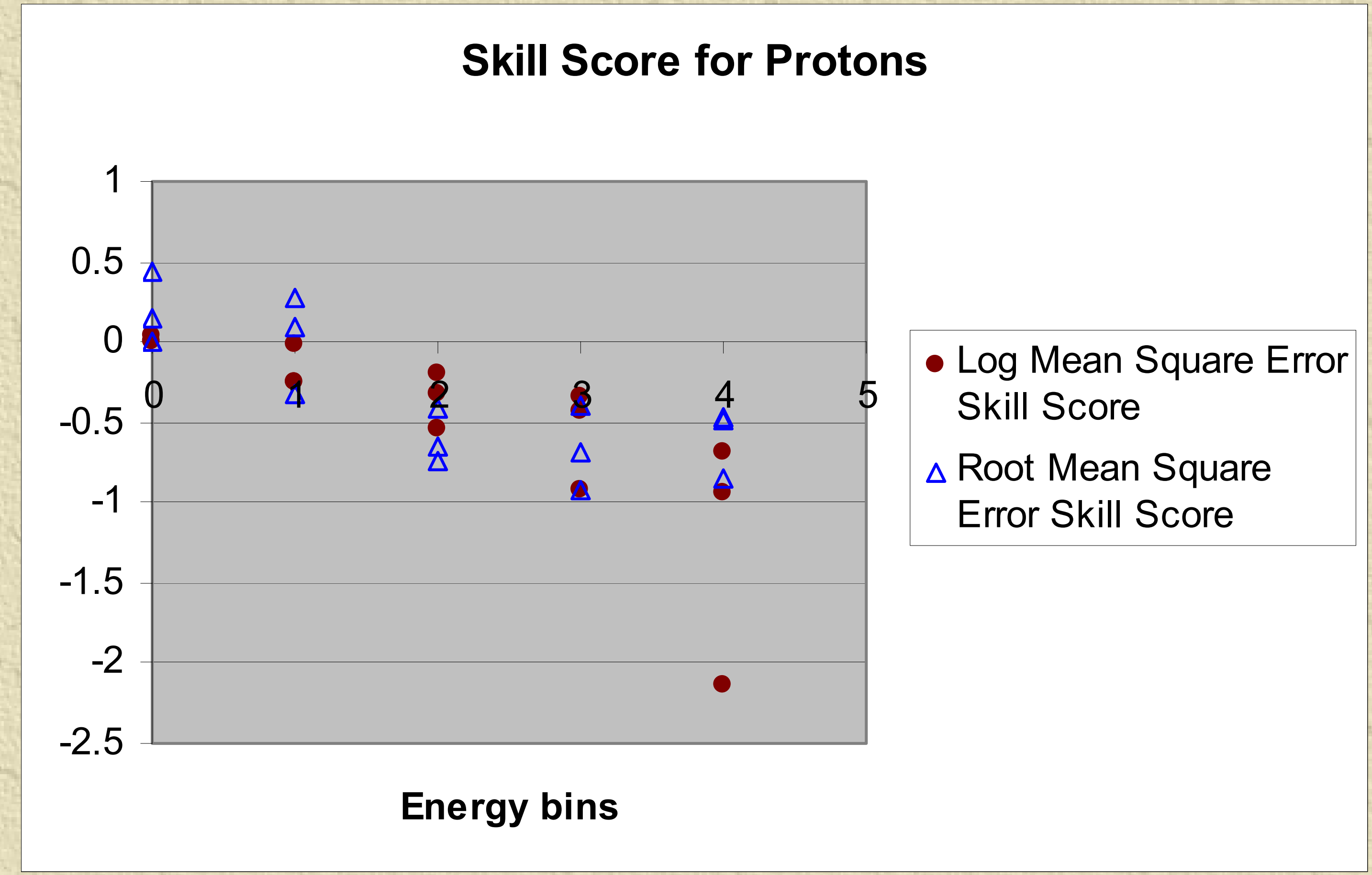


Energy Band (keV)	Log Mean Square	Root Mean Square
50-75	0.0203	-.995
250-400	0.0668	.232

Black is LANL data. Blue is the model results.

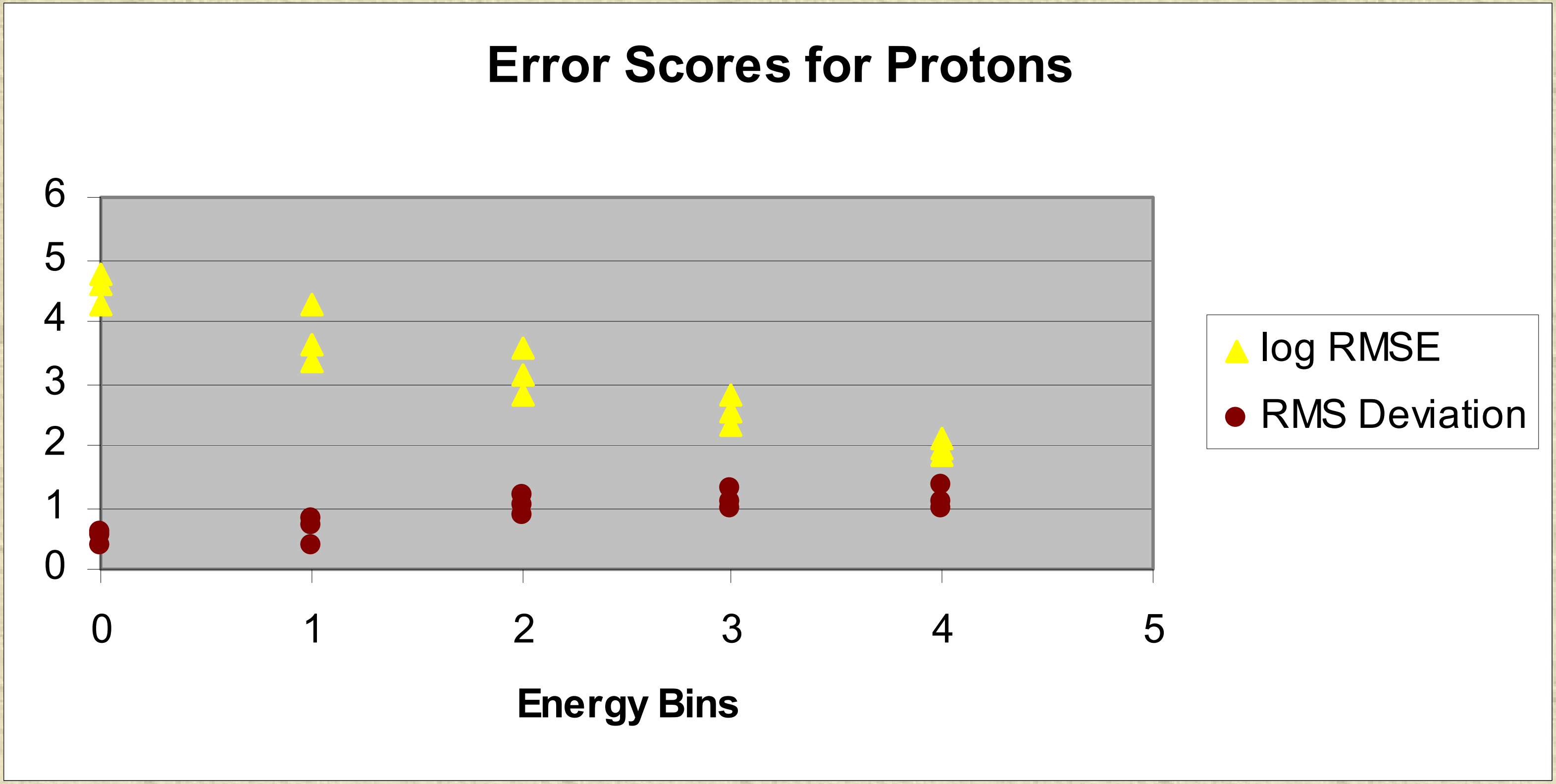
Geosynchronous proton flux data was provided by the Energetic Particle team at Los Alamos National Laboratory, Richard Belian (PI).

Skill Scores by Energy Bin Storm



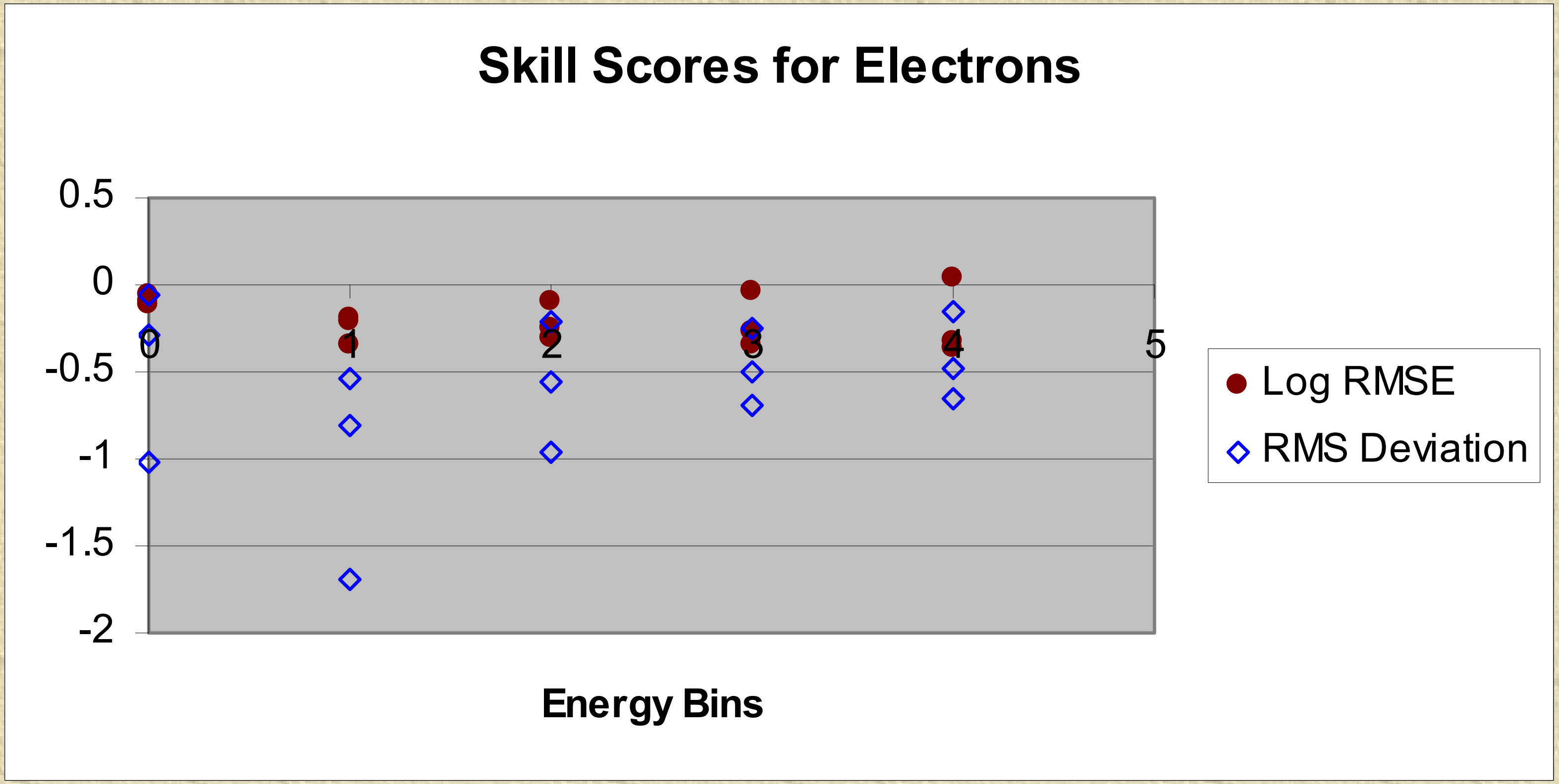
Energy bins (keV): 50-75, 75-113, 113-170, 170-250, 250-400

Error Scores by Energy Bins Storm



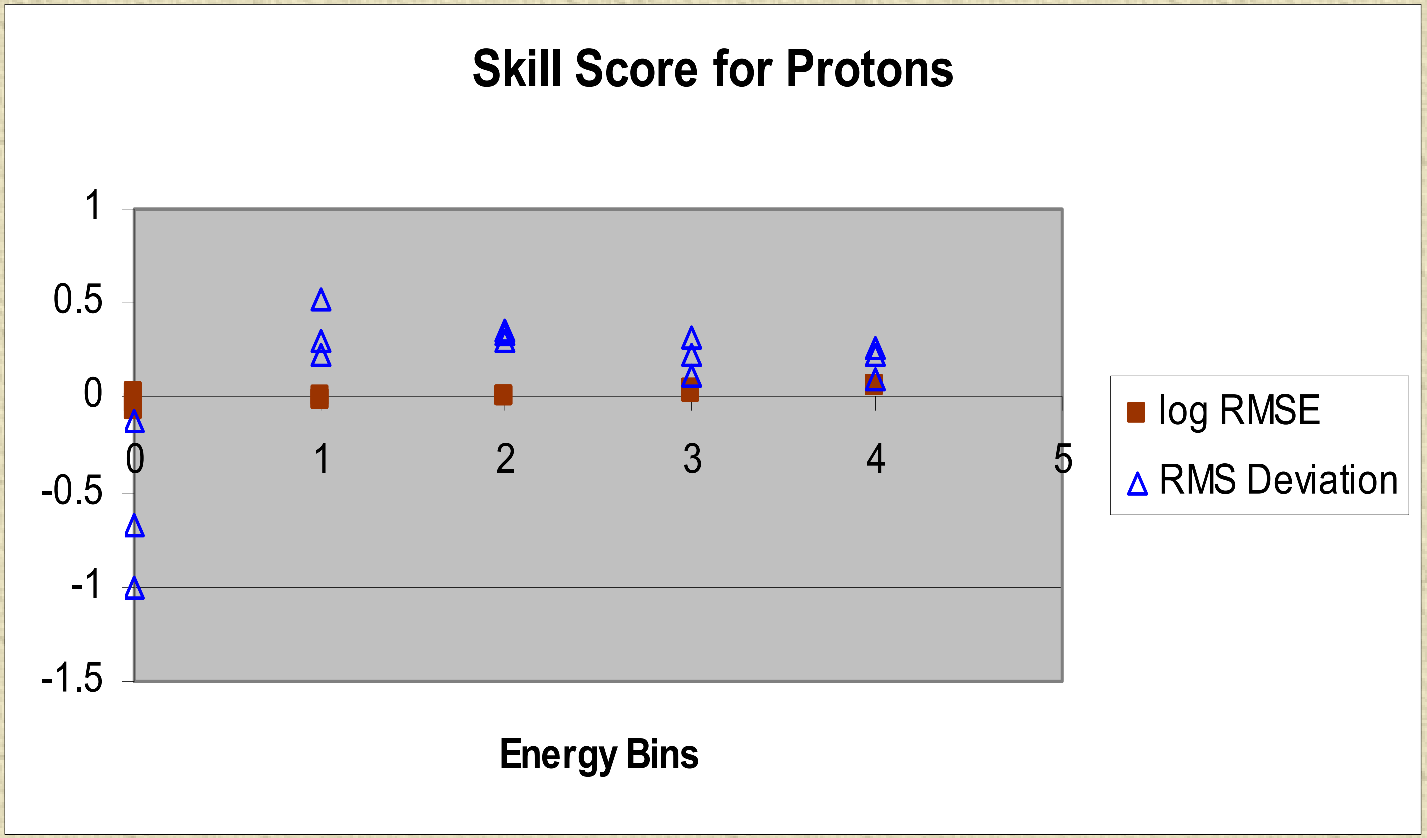
Energy bins (keV): 50-75, 75-113, 113-170, 170-250, 250-400

Skill Scores by Energy Bin Storm



Energy bins (keV): 50-75, 75-105, 105-150, 150-225, 225-315

Skill Scores by Energy Bin Sawtooth



Energy bins (keV): 50-75, 75-113, 113-170, 170-250, 250-400

Discussion

- The RMS deviation skill score tends to show a larger variation in scores than the Log RMSE skill score.
- On the sawtooth injection day, the model tended to get an average flux right but did not see any variation. The RMS deviation skill scores were high for four energy bins (around .2) while the Log RMSE tended to be around 0.
- The skill scores for the electrons were significantly lower.

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

Future Metric Domains

Need community input on metrics for

- Solar models
- Heliosphere models
- Ionosphere 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.

Proton Fluxes

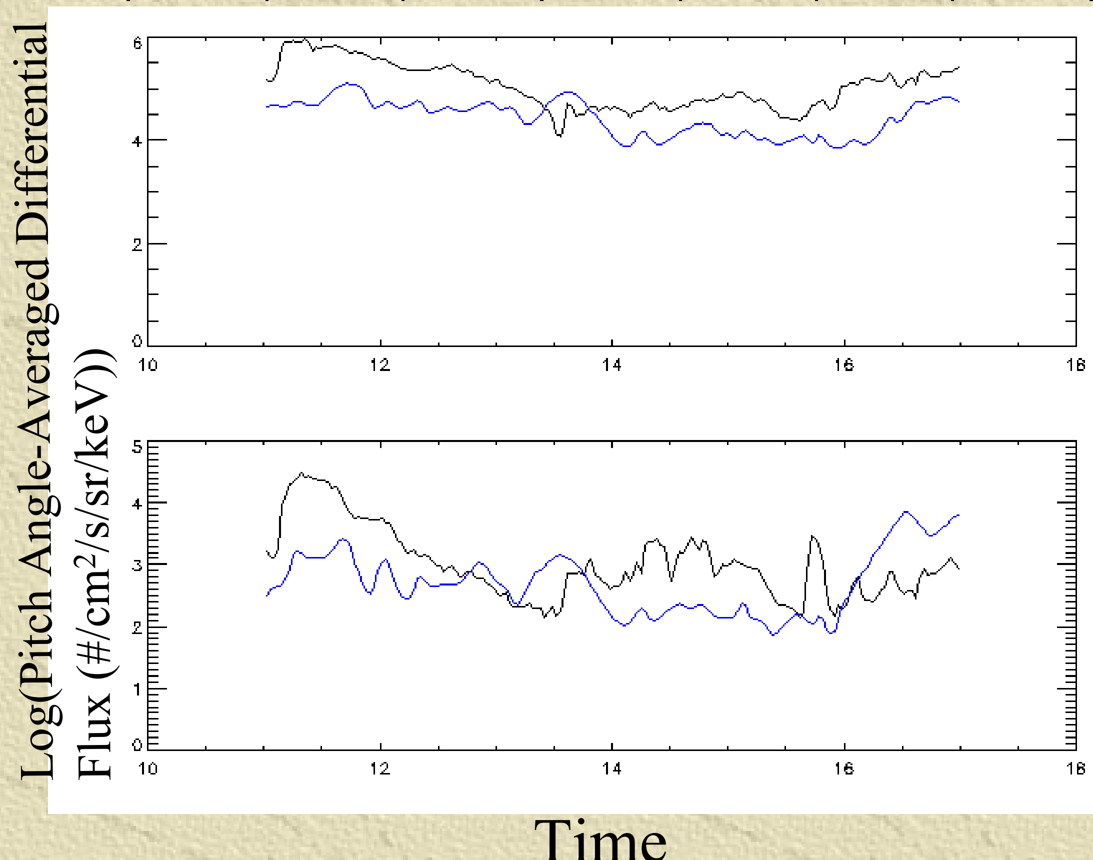
- Data
  - Proton fluxes from LANL geosynchronous satellites
- Model
  - Fok ring current model coupled to MHD models
- Root Mean Square Error Skill Score
  - Calculate root mean square error (RMSE)

    $RMSE = sqrt(\sum(predicted - observed)^2/npts)$
  - Calculate standard deviation of observations

    $STD = sqrt(\sum(observed - mean)^2/npts)$
  - RMSE skill score

    Skill score = 1- RMSE/STD
- Cross Correlation

Sample of Ring Current Metric



RMSE Skill Score	Cross Correlation
0.07	.59
-0.01	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).