The goal of this session is to discuss such failures, and assess whether they are due to limitations in observations, modeling, forecasting methods, or our understanding of the physics involved.

1st Scene Setter: KD Leka: Skill Scores and Evaluation Tools for SHINE-related Phenomena

- Verification is not just relevant to space weather - validation is healthy for model, theory, science in general
- Why validate? Substantiate claim, compare with others, learn from poor results.
- It is important to define to success/failure for your model – definition varies e.g. If a region has >50% probability of a flare, and two occur, is that a failure?
- Use appropriate skill scores to assess performance
- One skill score is not sufficient to get an idea of method is performing – different skill scores tell you different information.
Skill Scores and Evaluation Tools for SHINE-related Phenomena

• Why do research results fail to improve forecasts? Small sample sizes, event selectivity, data and computational requirements, statistical methods

• Small sample sizes lead to uncertainty, and statistically insignificant results

• Provide error bars for skill scores
Evaluating Success/Failure

Contributions and Discussion:

• Lan Jian: How reliable is the prediction of the solar wind background?
  – Comparison of many models for 7 solar rotations at Earth and Ulysses (higher latitude).
  – Various skill scores/assessment methods were considered together to verify performance.

• Keep in mind magnetograms (used as inputs to models) are themselves a model.

• Risk analysis: established framework in the disaster preparedness community, could be good to look into this for space weather. More on this today.
2nd scene-setter: Nariaki Nitta

- “What does space weather forecasting succeed?”
- Difficulties in determining flux-rope type from solar observations.
- EUV observations are not a substitute coronagraph data, for example some dimmings do not lead CMEs
- Multipoint observations like from STEREO are important to find the 3D CME direction.

Contributions and Discussion

- Robin Colaninno: GCS fitting of interacting CMEs is important to understand their evolution and can be important for Earth impact.
- Phil Hess: A separate fit for the CME and the shock/sheath together with a drag-based model can produce much more accurate results.
- Importance of having an L5 mission.
Discussion Topics for Today’s Session

- How do you define and quantify “failure” (or “success”)? (show us a one-slide example).
- What are appropriate (or inappropriate) “skill scores” for quantifying success? Why do "good" research results (e.g., that imply a reliable correlation) sometimes fail to improve forecasts?
- How do researchers benefit from transitioning their result into forecasting? How can a prediction environment help verify results and find the "ground truth."
- What are key differences between a forecast and a non-forecasting research product?
- What are examples of research results and methods that have led to a dramatic improvement in forecasting?
- What is the range of things being forecast? What aspects of these forecasts do we most want to have improved?
- What areas do we perceive to be particularly "weak" in our forecasting capability?
- What are the key obstacles in moving a principle into practice?
- What has been learned by going back to remodel failed events?
- Do we adequately incorporate the errors associated with observations when doing model validation/forecast verification/research?