What is required for successful Geoscience studies

I think the starting point is to articulate what we, or at least I would need and want in order to carry out scientific investigations. Then we can try to understand the best ways to create useful relations and connections between the physics and the computer science.

- I work in the ionosphere-thermosphere (IT) system. To study and understand the physics of the system, measurements or estimates of all the state variables of the system, globally, with spatial-temporal resolution required for the physics.
- In particular, we need to know neutral number densities and the corresponding ions, and the momentum/velocities and temperatures for each neutral and ion species.
- We need these on spatial scales ranging from global scales to meters, and on temporal scales ranging from decades to seconds.
- In addition to all of this, to close the system we need to know the external forcing, from the solar wind and magnetosphere, as well as from the mesosphere below, on similar spatial-temporal resolutions.

It is important to note, that for our goals scientifically, we are interested in understanding the physical processes and coupling between state variables, external forcing, boundaries between regions, and coupling across spatial and temporal scales. So, finding a correlation or relation between things, is just the start of the process. We always ask the question as to why do we see this relation or this response? What is the physical chain of processes that are causal?

What we have / issues / needs

Above, is a description of an ideal system. In reality, what we have is the following:

- A large number of different instruments that measure a number of different aspects of the system.
- Measurements are distributed globally, irregularly, in space and ground.
- Often measurements are someone else's noise, so we have no control.
- Researchers are distributed over a wide international community collecting their own data, doing their own processing, and making available to the rest of the community in an irregular haphazard fashion.
- For some state variables – such as electron density we have a rich set of observations. For others variables such as neutral winds and neutral composition the observations are sparse and hard to make.
- Our numerical models are often based on older numerical methods, are limited in spatial resolution, and rely on knowledge of the external forcing and boundary conditions.

Computer science community interactions

Based on the above, there are a lot of areas where there is an interface between the physics based research and computer science research fields.

- First, getting on the same understanding with language. When we hear things like: "information integration, machine learning, knowledge representation, geospatial computing, collaborative systems, visualization, and intelligent user interface", we don't know what that means within the context of our research areas. We need concrete examples that show one to one mapping that (for example) go between "information integration" and the research we do. Otherwise, we cannot really assess how much use it will be.

- One interface would be getting all these global heterogeneous data sources on one database / cloud / location for download. Typically we download about 10 different types of data over the globe for our data assimilation algorithms. We have scripts to go to all these different sites, get the data, process it, and put it in the correct directory structure. It would be very useful to have just one source. In addition, there are certainly a lot of data sources are not aware of. It would be useful to have an automated system that collects data from as many global data sets as possible, and put them in one place.
- A second interface would be defining standards and standard-processing methods to get all disparate data of the same type onto a uniform regular set of outputs. An example is magnetometer data: Our colleague, Jesper Gjerloev has spent a significant effort trying to make uniform the processing and output format for magnetometer data – called SuperMAG. While such a task might seem straightforward, my understanding is, it actually took multiple staff years to carry out. Having a central repository and a consistent set of processing tools to do this for all data sets would be great.
- A third interface of collaboration would modern computer methods at the boundaries of statistical estimation theory that help to estimate the state variables from measurements, fusions of models and data. For example, with very fast massive parallel systems, can we now apply cutting edge non-linear estimation problems such as particle filters? Our problem is a large phase space problem. We can easily have 1 Million grid points at each time step, and have 10's of state variables needed to be defined on each grid point. Then we need to evolve the system, and depending on scale, at few second time steps. Thus advanced numerical methods that can quickly sample such a large non-linear phase space, use what data is available to select the correct set of state variables at a given time, and then move forward in time would be a big step forward.
- A fourth collaboration, similar to the third is using cutting edge computational numerical methods on our IT numerical models. Most of our models use finite-difference methods, and do not take advantage of advanced methods designed for massively parallel systems. We would love to have self-consistent ionosphere-thermosphere coupled models with electro-dynamics that have all the non-linear terms and go down to small spatial-temporal scales. This will naturally lead to questions of chaos and turbulence.
- Another area is intelligent systems that can take recent measurements – for example from the incoming solar wind and/or high latitude response, and try to predict what the forcing will be several hours into the future. Since we are so strongly forced, any predictive model can have all the state variables known exactly at an initial time, and the forcing will completely de-correlate the solution on a scale of a few hours. Since it will be a long time before we can measure the solar wind stream at the Sun, accurately predict it at Earth, then accurately model the magnetospheric response, storage, and eventual release of energy to the IT system, we really need smart intelligence-based empirical methods.