Red Group Notes:

Reporter: Matt Mayernik

Q1 – What have you heard that most excites you

- What Cliff Mass talked about, probabilistic forecasting and what he has done in public interaction with those tools and services. Developing an infrastructure where the public can access prob. information is a good thing. Allowing them to see what can be done there.
- Trying to bring the data together is very good thing, linking between data sets is important, also need easy to use formats. Need to have data and interfaces together. Large data sets themselves can be hard to use, interfaces and visualizations are critical components. I.e. ways to browse and view the data before writing scripts to work with the data.
    - Seconded by another member
- Took a while to try to understand what the workshop is leading to. Seems that this points to seeing the three pieces together, data, models, and tools (like Unidata tools).
- Potential to bring all the data together in a central repository is very exciting, particularly if they can be linked. Also a central place to find software. But have some doubts about how it can work out.
- First thing we need to do is build a community, not just build up a data pool. Need to learn what is the best and most useful information first. Want to know quality of data, in addition to data.
- Some central access layer where data can be accessed with good metadata.
- That information about the data could be more available. That data could be more findable and usable.

Q2 – What are the key challenges for advancing DA and Ens. forecasting

- Many of the biggest challenges are sociological not technical.
    - [Most people in the breakout agree]
    - People are set in their ways and comfortable in what they do, need to get agreed-upon metadata
- Need ways to motivate people to put data, or link, to the central location.
    - Even if links available, also a challenge to produce user friendly data.

    - Might find data but people don't answer questions about them.
    - If resources are tight, you don't want to be doing end-user support.
- Lack of resources, man power and funding.
    - You might know what to do, but not have enough people and funding to do what needs to be done.
    - Who is going to do this? This is a key question.
    - Will NSF support activity like making data available for secondary use, documentation, user support, etc.
- Being able to get through computational aspects to be able to answer actual questions. Time reduction is desired.
- Teaching – interfaces or ways to allow students to get more into data.
- Linking repositories together through some kind of portal or central interface.
    - Would be nice if this central portal also included publicly understandable
    - When you have central repositories, currently, unless you are part of an inner crowd, you don't know that big repositories exist. Google and Bing, etc, can find metadata if the catalogs are exposed. But that is lacking: providing rich metadata and exposing the metadata to search engines so that they can be discovered.
    - Need to thinking about how to expose the catalogs to the largest possible audiences, including search engines.
- Model techniques have limitations, maybe have devoted too much into DA and ensemble forecasting instead of model development and quality assessment.

3. What are science drivers?

- Understanding what ensemble forecasting should be about. Understanding non-linear flows, deterministic vs. probabilistic predictions. How much of the ensembles are numerical vs. related to physical principles? Linear modeling techniques might not directly provide good results for non-linear systems. Hybrid linear and non-linear systems are not properly evaluated through ensembles right now. Ensembles might not be close to accurate. Evaluation is also largely based on linear methods, not incorporating non-linear structures, like fractals. We haven't evaluated what non-linear flow should look like.
- NRC report talks about changing our language/approach from getting weather forecasts to getting impact forecasts. Need more data, in order to assess what impacts are (and what "impacts" mean in the first place)

4. What infrastructure advances are needed?

- Connecting tools like IDV to disparate data sets in an online fashion, so that you could run IDV remotely without downloading the data.

    - Making data sets analyzeable online, so that they can be viewed and analyzed where they are, on the server where they are hosted.
    - Need to be sensitive to internet speeds. Need to develop efficiency with regard to variable internet speeds.
- Easier way to access data in fewer formats. Students struggle with too many formats. Would be nice to have a search interface for particular kinds of data products, in particular formats, perhaps uniform formats or ways to convert among formats
- Not clear that appropriate metadata standard for observational data exists. It would need to much more widely used than we have now.
- Good to make requirements for how data provider must provide to data repository. I.e. set up standard for people to follow.
    - Problem is that with many data sets, we will have no leverage on data providers, so maybe it could be levels of providers, such as Gold level (highest level of quality and support, etc) down to low level of standardization.
    - "Long tail" scientists don't have resources to do high level of data and metadata quality control and standardization. Perhaps EarthCube could provide tools or services to do translations, standardization, etc. ACADIS as an example: individual polar researchers who commonly use excel spreadsheets, no conventions, lots of diversity, but there is some awareness on part of PIs that this is problem and that some central way to do it is necessary.
    - Can there be some facility that helps individual PIs to take data or software and turn it into community data or tools? Something like software institutes.
- Cannot separate data and software, need ways of keeping the two connected.
- Hardware is a huge bottleneck, e.g. quotas on disk space that you can use, which means that things always get thrown away, which limits advances. Need to have storage advances. Forced to throw stuff away because of hardware limitations
    - Problem now exists that data generation tools are outstripping storage growth rates.
    - But what might be addressable is compression, filtering, subsetting approaches.
- Some high profile field projects have addressed some of these issues [discussed above] because resources have been put towards addressing them.

5. What is appropriate level of community coordination and governance structure is needed to facilitate all of above.

- One possible approach would be to base the organization on the different science groups.
- Working groups
- When you write a proposal, you write a data management plan. That is the place to start. NSF could provide follow-up support or guidelines on what to do once funded. Same with software sharing plan. E.g. downloading NASA data, might not need to share the data, but could share the software.
- Ex. Conduit project – brought NCEP data to universities. It worked with NCEP, data providers, and the university to come to agreements about what data to bring in, etc.
- EarthCube is many communities, so bringing them together is a challenge.
- ESIP might be a good template for this. A loose federation of groups, but centralized meetings, tools, and testbeds.
- Many building blocks exist. Are we giving more funding to particular sub-group to do coordination. A body to do this.
- The organization structure will be related to the funding structure.
- Important to be cautious. Just getting the data at high fidelities could be all-consuming. So it needs to be balanced with the science needs. Don't want to turn the NSF into a data organization, it shouldn't impede the core NSF research mission.
- Nobody coordinating across NSF projects right now, i.e. Unidata, DART, etc. What would you call this organization? Coordination should get specific funding.
- Starting point is where we are now, coordinating existing organizations/projects. Big centers might be better positioned to do this because, a big jump forward would be hard for university people to do. But there needs to be university-based steering and prioritization.

Q6 – What do you see as broader impacts of EarthCube

- Multi-disciplinary collaboration
    - Seconded, right now, we often fail to bring together related studies.
- More students want to get into data assimilation, answer science questions that now they might get scared away from due to technical difficulties.
    - Seconded, students can be scared away by technical requirements
    - Thirded, need to have students on the front edge of techniques like probabilistic forecasting
- Pushing data and science results out to more communities (e.g. public and education)

Q7 – What do we see as the next steps

- Need to reach out other communities, like air quality and hydrology.
- Need to have more internal discussion related to DA, etc, and about how to integrate with other research goals.
- Develop a prototype system that links data sets together, involving the most used projects, like maybe reanalysis data sets. Develop a system that works. From that branch out to broader cross section of data sets.
    - This could help us think about science drivers.

Q8 – What have we missed?

- Other atmospheric science communities. Also, nobody here from NASA, EPA, NOAA, other organizations.
- Non linear mathematicians.
- Social community network, would help to get people involve. Emails and surveys are limited in response, but people might be more inclined to post on social media. Might facilitate communication.
- Metadata was discussed in passing in many talks, but not explicitly as much as might be useful.
    - Related – data provenance
    - Data publication – citing data sets as a way to provide incentives.
- Data mining experts could have a lot to contribute.