Some of the most interesting properties of the climate system are emergent (e.g., sensitivity to external forcings, predictability at the regional scale). By emergent we mean a property that arises from complex interactions between, for instance, dynamics, radiation, cloud formation, and surface fluxes, rather than being a function of a single physical process. Most of the traditional global-scale diagnostics used for climate model evaluation are similarly emergent. Emergence therefore complicates our ability to attribute a specific piece of code or model assumption. Indeed, model developers are often left to their experience and trial-and-error when addressing these discrepancies. Unsurprisingly, some notable discrepancies have persisted across multiple generations of climate model development (e.g., the double ITTC problem). Even with the availability of large archives of coupled GCM output (e.g., CMIP5) and complementary observations to go with them (e.g., Obs4MIP) our ability to address certain questions is limited.

There are three main sources of model/observation discrepancy: 1) the model is deficient, 2) the data is deficient, or 3) the comparison is inappropriate or misleading. Many things lead to the third situation. For example, Eulerian time averaging is often seen as a desirable form of data compression for such comparisons. However, the concept of emergence means that each field is strongly influenced by the cumulative actions of intermittent and transient phenomena that cannot be seen directly in the time mean of the field (e.g., convective storms and cyclones). As a result, comparisons using time means are unlikely to reveal why a discrepancy exists (i.e., situation 3). This suggests the need for a more effective approach to diagnostic-based model development.

Often these approaches can take the form of a Lagrangian conditional average, which when done correctly, merges a case-by-case perspective of single events with the statistical approach required by climatologists. In this way process-based diagnostics (PBDs) broaden the pool of traditional climate model validation methods.

**Introduction**

**Use Case - Extratropical Cyclones**

Extra-tropical cyclones make excellent candidates for PBDs because: 1) Cyclones are specific, identifiable and well-understood phenomena. 2) Cyclone activity shapes the distribution many quantities on both climatic and weather scales (e.g., cloud, temperature, wind). 3) Cyclones have interesting internal and external variability. 4) While today’s climate models can in principle resolve basic cyclone features, they are less able to represent smaller key features (e.g., fronts), and questions remain about their ability to capture more subtle changes in cyclone behavior and structure (e.g., variations between seasons, hemispheres). Indeed, mid-latitude cyclone clouds are a key source of inter-model difference in climate sensitivity (Williams and Tselioudis).

An ongoing project led by one of us, “The MAP Climatology of Mid-latitude Storminess” or MCMS, is designed to address just these sort of questions (see Fig. 2, http://gcss-dime.giss.nasa.gov/mcms/mcms.html).

Here we compare the SLP fields from the NCEP Reanalysis II (NRA2) and a climate model (GISS-E2-R) run with complementary historical boundary conditions for the years 1990-2010 (21 years). Fig. 3 depicts the traditional approach of examining the time-mean fields, which in this case are generally similar except that the GISS result is systematically lower pressure especially over the ocean.

**Use Case - Extratropical Cyclones**

**PROBE**

**Automated Event Service (AES)**

AES is a NASA funded project that is developing a tool for systematically finding events of user-defined Earth science phenomena in large collections of data. AES provides links to additional remote sensing imagery which overlap with the discovered event tracks. Initial work uses regularly-gridded data such as multi-decadal reanalyses (e.g., MERRA), but will eventually be extended to include remote sensing data. Examples of events include:

- High snow over Europe
- Tropical cyclones
- Tornadoes
- Urban heat islands
- El Niño/ La Niña
- Above average temperature across Antarctica
- Heat waves
- Dust storms
- Monsoon jets
- Blows snow over polar ices
- Sea ice extent
- Ozone abundances
- Dust plume activity
- Aerosols
- Sea ice
- Wind fields
- Flood hazards
- Watershed changes
- Forest fires
- Water vapor
- Surface temperature
- Water temperature
- Clouds

**ScDB**

ScDB is built on top of SciDB, a massively parallel, open-source, next-generation database technology that has strong support for multi-dimensional data and provenance. The Array Functional Language allows users to extend and customize the built-in analysis capabilities.

**CCL Extension to SciDB**

Connected Component Labeling (CCL) is a vital capability for PROBE. Because this algorithm could not be expressed in a purely data-parallel manner, efficient implementation within SciDB required the creation of a custom User-Defined Operator (UDO) for 4-D spatiotemporal data.

**Workflow Automation**

Aside from implementing various custom analysis procedures such as CCL, PROBE will primarily extend AES by incorporating workflow technology to automate the process of creating PBDs. Such automation will:

- Re-apply PBD analysis to auxiliary datasets
- Automatically ingest new model output to SciDB
- Recompute previous PBDs as new data arrives
- Notify user when new results are available

**References**

- Bauer, M.: The MAP Climatology of Mid-latitude Storminess (MCMS) http://gcss-dime.giss.nasa.gov/mcms/mcms.html
- SciDB (http://scidb.org)