Next Generation Climate Architecture

Introduction

The purpose of this document is to describe a next generation climate architecture based on the comparison of climate models and remote sensing data that are used to inform regional decision makers, states, and federal government and (inter-)national stakeholders who make critical policy decisions involving the weather, future climate, state and regional level tourism, water resources management, food management and security, etc., based on this information. The "Next Generation Climate Architecture" combines both modeled simulation output of current, historical and future climate scenarios alongside remotely sensed observations acquired from the National Aeronautics and Space Administration (NASA), the National Oceanographic and Atmospheric Administration (NOAA), EPA, USGS, and other sources both from space and sub-orbital spatial coverage.

Motivation and Background: Climate Models and Measurements

Future estimated projections of the Earth's climate as derived from *Climate Models* suggest drastic changes in individual and combined measured parameters such as temperature, for a variety of reasons including greenhouse gases and other human influenced parameters[1]. Climate models are physical and heuristically based models of Earth system dynamics involving the ocean, the atmosphere, land, and other specialty domains. Climate models are traditionally developed by specific modeling groups whose expertise may reside in one or more of these domains e.g., at Institution X they specialize in the development of ocean models, whereas

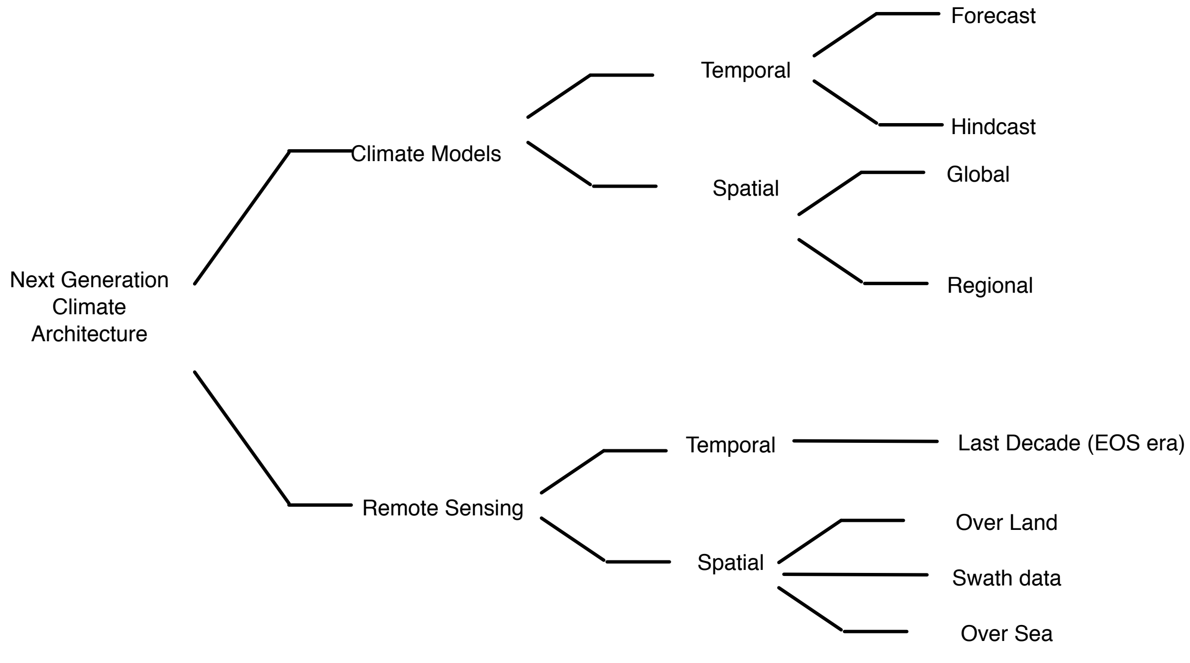

Figure 1. Climate model and remote sensing data considerations (spatial + temporal)

[1] http://www.epa.gov/climatechange/science/future.html

Institution Y is excellent at atmospheric models and also land surface models. Climate models simulate one or more *Parameters* that are dimensions of some Earth domain, e.g., for the atmosphere, parameters may include *Temperature*, *Solar Radiation*, *Heat Flux*, etc. or for the ocean, parameters may include *Sea Surface Temperature* or *Wind Speed at the Sea Surface* or *Salinity* etc.

Climate models traditionally have a temporal span and nature, they may cover e.g., a decadal time span such as 2000-2010, or they may be a 50- or 100-year span simulating into the future (*projection* or *forecast*) or they may simulate the past (*hindcast*) observed environment parameters. Furthermore, climate models also traditional have a spatial coverage, they may be global and cover the Earth at some NxM degree grid box, uniformly spaced, or not; or they may be regional, e.g., the Western US, and be traditionally at much smaller and more precise resolutions.

The ability of climate models to simulate and output parameters of physically derived quantities like temperature or wind speed relies very much on our ability to *observe* and measure those values and to provide some basis on which to derive the mathematics and physics needed to perform these simulations and thus to produce *climate model outputs*. Observations typically come from different ground sources such as stations and towers, or in-situ (based on hand held instruments); they come from airborne platforms (helicopters, jets, etc.); and they come from space borne missions such as those that are flown by NASA, and NOAA, etc. Data acquired from non-ground based or in-situ data is *remotely sensed* or *remote sensing* data.

The various dimensions of dealing with climate models and remote sensing data are high-lighted in Figure 1.

Accessing, Analyzing and Making use of Climate Information: Why so Difficult?

As previously stated climate models (their outputs), and climate observations are generated from a variety of sources, and by a variety of institutions including governmental agencies like NASA, NOAA, EPA, etc. Because of the geographically distributed nature of these institutions, and because of the heterogeneity of the sources of climate information, bringing together for example the climate models (their outputs) and the remote sensing observations to compare their measurements of the same parameter is quite difficult. Consider the case in which a climate model simulates *Sea Surface Salinity* and there is a similar NASA remote sensing mission that deploys an instrument that also observes this parameter from space. The remote sensing data may be stored in the Jet Propulsion Laboratory's Physical Oceanography Distributed Active Archive Center (PO.DAAC), and may be available via FTP as a Hierarchical Data Format (HDF) version 5 file, with associated HDF-EOS metadata (or "data about the data") describing where the data was captured (its temporal and spatial bounds), information about the mission, etc. The climate model output may be available through the HTTP/REST OPeNDAP protocol from the Earth System Grid Federation (ESGF) and one of its many replicated nodes throughout the US and the world, as a NetCDF formatted file with Climate Forecast Conventions (CF) metadata.

Note the variations in *data file format* (HDF versus NetCDF), *metadata* (HDF-EOS versus CF), *protocol* (FTP versus OPeNDAP), not to mention the other differences that would have to be mitigated to effectively compare the same measured parameter. For example the climate model output for *Sea Surface Salinity* may have a temporal range of 2000-2010, but the NASA remote sensing data may only begin starting in 2008; also consider that the remote sensing data may only have discrete values for the times that the instrument takes data (e.g., 1am and 1pm). On the spatial context, the climate model may produce global estimates of *Sea Surface Salinity* whereas the NASA remote sensing instrument may only take data in a swath geometric pattern with non uniform grid cells spatially.

Various statistical means are available to mitigate these differences. For example we could compute an interpolation or average of the discrete time measurements to allow interrogation of the remote sensing data at any time (like the climate model output provides). We could also spatially interpolate or average the remote sensing data into uniform grid cells like the climate model output. Any of these strategies are both *computationally intensive* as well as *data intensive* considering the large amount of information (decades worth) and precise spatial resolution that each measured or simulated value of *Sea Surface Salinity* provides.

Once the values from the model output and remote sensing data are comparable, we may then compute some distance measurement or *metric* that allows their comparison such as Bias, or Root Mean Squared Error (RMSE) or we may compute a probability distribution function (PDF), etc. These are also *data* and *computationally* intensive operations.

Finally, with a computed distance metric between the model output and remote sensing observation, we can visualize the difference using one of many toolkits such as the NCAR NCL command language and visualization package; or by using Matplotlib from Python, or by using Matlab or R to demonstrate this variance. These visualized comparisons are typically what is fed into decision makers hands and what allows prediction of simulated versus observed measurements and what informs policy decisions based on the climate.

The Next Generation Climate Architecture

A system that we will call the "Next Generation Climate Architecture" is being developed that will deal with these challenges and requirements that we have outlined. Consider the following rough "Level 0" architecture of such a system demonstrated in Figure 2.

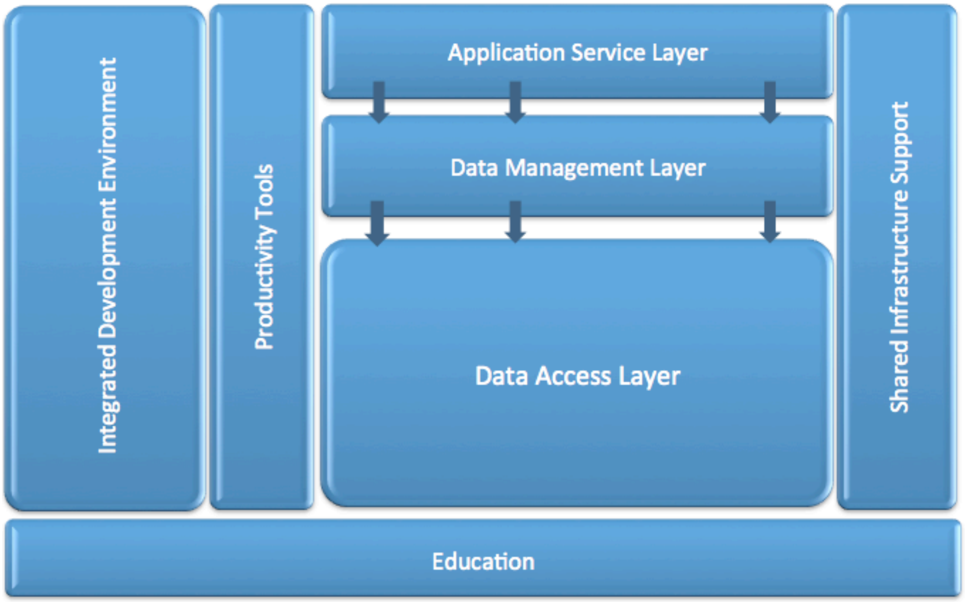
Integrated Development Environment
Productivity Tools
Application Service Layer
Data Management Layer
Data Access Layer
Shared Infrastructure Support
Education

Figure 2. Next Generation Climate Architecture Level 0

The architecture depicts several components that we will briefly describe in this paragraph. The *Integrated Development Environment (IDE)* in the left side of Figure 2 is responsible for allowing users of climate model outputs and remote sensing data to design applications and systems that interact with the data, that perform regridding and metrics computation, etc. The *Productivity Tools* component defines mechanisms for notifying users when Climate Model outputs and remote sensing data is available potentially through publish and subscribe mechanisms; or by allowing users to request these data on demand, or more information about them on demand. Productivity tools may also include decision support applications that make use of climate information. The *Application Service Layer* in the top middle portion of Figure 2 defines core services for apps to be constructed that process, acquire, visualize and transform climate model output. This could involve the definition of interpolation and regridding services; and may involve statistics and physical models and for the definition of metrics to compare model output with observations. The *Data Management Layer* provides functionality for registration (cataloging and archiving) of remote sensing data and climate model output; and for basic workflow management and processing of these data (e.g., transforming their formats; accessing and manipulating their metadata, etc.). The *Data Access* layer is responsible for locating and for providing climate model outputs and remote sensing data from distributed sites, mitigating security and institutional policies as needed. *Shared Infrastructure Support* provides for allocating data processing jobs to compute resources be them cloud resources (e.g., Amazon); shared grid resources across institutions; or local computing (e.g., laptops, desktops, etc.). Finally the *Education* component is training, documentation and architectural information allowing others to understand the next generation climate architecture, its styles, components, etc.