Validation and quantification of uncertainty in coupled climate models using network analysis

Grant: DE-SC0007143

Final Technical Report

Period Covered: October 2011 to June 2015

PI: Annalisa Bracco,
School of Earth and Atmospheric Sciences
Georgia Institute of Technology

Co-PI: Constantine Dovrolis

Funded by:
Climate and Environmental Sciences Division
of the U.S. Department of Energy
Office of Biological and Environmental Research (BER) program

U.S. DEPARTMENT OF ENERGY | Office of Science

Table of Contents

(A) Project Overview

General circulation models currently used for understanding current and past climate and for predicting its evolution in the future, exhibit substantial spreads in their equilibrium sensitivity, implying that the magnitude of their temperature increase in response to a doubling carbon dioxide is uncertain. The mean temperature increase over the 21st century projected by models in the last Intergovernmental Panel on Climate Change assessment continues to resist any narrowing of the range of estimates even in the historical integrations while the evolution of the major modes of variability of the climate system diverges.

Large uncertainties for end-of-century climatic variables prevails not only in the simulation of future surface-air temperatures, but also in precipitation, cloud cover, winds, sea level, sea ice, and other variables of importance for socio-economic, ecological and human-health impacts. It is fair to state that while no legitimate doubts exist about the future rise in global temperatures and about additional changes in climate being significant, many questions remain about the extent of the changes, not only in the mean, but also and even more so in the variability of climatic fields, in space and time.

The fast growing availability of observations from remote measuring platforms such as satellite and radars, as well as the increasingly more detailed outputs from global-scale climate models, contribute a continuous flow of terabytes of spatiotemporal data. The last two decades have been characterized by a rate of data generation and storage that far exceeds the rate of data analyses. New approaches to quantify and characterize uncertainties in climate model simulations are required, and some have been developed in recent years. While the literature in statistical analysis applied to climate fields, observed or modeled, is mature, systematic efforts in climate data mining are still lacking.

The first research goal of this project was to develop a novel, fast, scalable and cutting-edge computational toolbox based on complex network analysis to investigate local and non-local statistical interrelationships in climate model outputs. In computer science, “complex network analysis” refers to a powerful tool used to investigate local and non-local statistical interrelationships. Such tool is composed by a set of metrics, models and algorithms commonly used in the study of complex nonlinear dynamical systems, and its main premise is that the underlying topology or network structure of a system has a strong impact on its dynamics and evolution.

The second research goal was to apply such method to the Coupled Model Intercomparison Project Phase 5 (CMIP5) to identify teleconnections and quantify how well they are represented in the historical period, and to evaluate their evolution in the future, under global warming scenarios.

(B) Network Analysis Framework

Since 2004 climate networks have been used to investigate climate shifts relating network changes to El Niño Southern Oscillation (ENSO) activity, identify global scale structures responsible for energy transfer through the ocean, evaluate climate models and identify teleconnections, and represent the interaction between different climate variables as a network. In most cases, edges between nodes of the climate network are inferred using linear or non-linear similarity measures (for example Pearson correlation, mutual information, or phase synchronization), and the network is constructed as a (weighted or binary) undirected graph. It is well noted that correlation does not imply causation and the next challenge in climate network analysis is arguably to move from undirected correlation based networks to directed causal ones to be able to identify feedback loops between the different variables of the climate system. Additionally, the network inference methods adopted in the works just cited construct graphs in which two cells are not considered connected and they are consequently pruned - or their correlations are set to zero - whenever the cross-correlation between them is less than a given threshold. Cell-level pruning makes the network inference process less robust and limits dramatically the reliability of model intercomparison exercises that implement it.

To overcome these limitations, to quantify differences and analogies between models or modeled and observed quantities, and to extend the application of network analysis to model ranking and intercomparison, in this project we have developed a novel, fast, robust and scalable methodology to examine, quantify, and visualize climate patterns and their relationships.

(C) Executive Summary

The project contributed so far to two peer-review publications, a book chapter, and two manuscripts in the latest stage of preparation. It supported a graduate student, Ilias Foundalis, through his graduate work (Mr. Fountalis expected date of defense is Spring 2016).
Below is an executive summary for the findings related to goals #1 and #2.

1) We developed a fast, robust and scalable methodology to examine, quantify, and visualize climate patterns and their relationships. It is based on a set of

notions, algorithms and metrics used in the study of graphs, referred to as complex network analysis. This approach can be applied to explain known climate phenomena in terms of an underlying network structure and to uncover regional and global linkages in the climate system, while comparing general circulation models outputs with observations. The proposed method is based on a two-layer network representation, and is substantially new within the available network methodologies developed for climate studies. At the first layer, gridded climate data are used to identify "areas", i.e., geographical regions that are highly homogeneous in terms of the given climate variable. At the second layer, the identified areas are interconnected with links of varying strength, forming a global climate network. The robustness of the method (i.e. the ability to separate between topological distinct fields, while identifying correctly similarities) has been extensively tested. It has been proved that it provides a reliable, fast framework for comparing and ranking the ability of climate models of reproducing observed climate patterns and their connectivity.

We further developed the methodology to account for lags in the connectivity between climate patterns and refined our area identification algorithm to account for autocorrelation in the data.

2) We applied the new methodology based on complex network analysis to state-of-the-art climate model simulations to assess their performances, quantify uncertainties, and uncover changes in global linkages between past and future projections. Network properties of modeled sea surface temperature and precipitation over 1956–2005 have been constrained towards observations or reanalyses, and their differences quantified using two metrics. Projected changes from 2051 to 2300 under the scenario with the highest representative and extended concentration pathways (RCP8.5 and ECP8.5) have then been determined. The network of models capable of reproducing well major climate modes in the recent past, changes little during this century. In contrast, among those models the uncertainties in the projections after 2100 remain substantial, and primarily associated with divergences in the representation of the modes of variability, particularly of the El Niño Southern Oscillation (ENSO), and their connectivity, and therefore with their intrinsic predictability, more so than with differences in the mean state evolution.

Additionally, we evaluated the relation between the size and the 'strength' of the area identified by the network analysis as corresponding to ENSO.

(D) Report of Findings

- Methodology Development

The network inference developed for this project is a three-step process. First we construct a ``cell-level network"; second we apply a clustering algorithm to identify the nodes or areas, i.e. non-overlapping geographically connected regions that are homogeneous to the underlying variable; third we compute weighted links between areas to assess their connections. The cell-level network is constructed computing the Pearson cross-correlation between the detrended time series of the climate variable of interest for all grid cells pairs. Quite naturally time lags can also be taken into account in the cross-correlation calculation to build a dynamical network. All pair correlations are retained and the resulting cell-level network is a complete weighted graph (i.e. a link exists between all pairs of grid cells). This characteristic differentiates our method from most prior work on climate networks where a threshold to prune non-significant correlations is applied, and ensures robustness of the area-level structure, allowing for reliable comparison of different networks, as extensively tested in the first publication that resulted from our effort (Fountalis et al. 2014).

The clustering algorithm relies on a single parameter, τ, that varies between models or datasets considered and controls the homogeneity of areas to the underlying climate variable. τ represents the minimum average pair-wise correlation between cells of the same area at a given significance level. The algorithm aims also to minimize the number of areas identified; the problem is shown to be NP-Complete, thus the algorithm must rely on “greedy heuristics”. Finally, links are computed from the area cumulative anomalies weighted by the cell sizes. The weighted link between two areas is equal to the covariance between the corresponding cumulative anomalies; links - positive or negative - are computed for all pairs of areas to obtain a complete weighted graph. Link maps allow the visualization of the (weighted) connections between any given area and all others in the network. Areas are also characterized by their weighted degree or strength, defined as the sum of the absolute link weights. Strongest areas exert the greatest impact on climate variability.

Using complex network analysis to evaluate models' performance and their dependencies yields several desirable properties. The investigation is not locked into a particular climate mode or index - or to a set of indices - from the outset. From a set of climate model runs, different users can evaluate networks for various fields and/or regions, and derive model-dependent areas and their links in lieu of climate modes and their teleconnections. The methodology is scalable, and allows for direct, robust comparisons between different models or the same model integrated using different parameters, parameterizations, or forcings. Furthermore, it is immediate to include an estimate of internal variability when multiple ensemble members are available, which can be directly compared to contributions from different forcings, and to model trajectories over time.

To quantify similarities and differences between two networks in a compact way, we developed a new metric and adopted one from the complex network literature. For example, if we consider networks N and N', for the same variable (for example sea surface temperatures for a realizations of the Community Climate System Model Version 4, CCSM4, and for HadISST, each of size n grid cells. First, their strength distributions can be compared using a newly defined network distance D. The smaller the distance, the more similar two networks are in their strength distribution. Additionally, the spatial likeness of the areas in the two networks can be quantified using the Adjusted Rand Index (ARI). Any pair of cells that belong to the same area in both N and N', or that belong to different areas in both networks, contributes positively to the ARI. Conversely, any pair of cells that belong to a given area in one partition but to different areas in the other, contributes negatively. The ARI ranges between 0 and 1, with 1 denoting perfect similarity. The metric is adjusted to ensure that the distance between two random partitions is zero. ARI and D can be considered globally – i.e. spaced averaged over the whole model domain - for example to compare the evolution of the network for a specific field under increases greenhouse gas forcing, or regionally –i.e. spaced averaged over a specific region -, to analyze the response of a limited number of areas.

The heuristic described above does not account for the possibility of not independent time-series in the process of selecting the area identification threshold (i.e. it is not appropriate if autocorrelations exist in the data). It is important to remedy to this limitation whenever time-lagged correlations are the focus of the investigation. A more appropriate statistic, accounting for the existence of autocorrelations in the data, is provided by the *Barlett* formula. We further developed our network methodology to include for it.

- Application to Reanalysis Datasets

We first applied the new methodology to a set of reanalysis sea surface temperature and precipitation data sets. An example of strength and link maps is provided in Figure 1 for the Hadley Center sea surface temperature (HadISST) reanalysis shown here for boreal summer (JJA) over 1956-2005. The strongest area identified in the network corresponds to ENSO and it is linked to the Indian Ocean where SSTs are found to be warmer than average in correspondence of El Niño events, and vice versa for La Niñas.

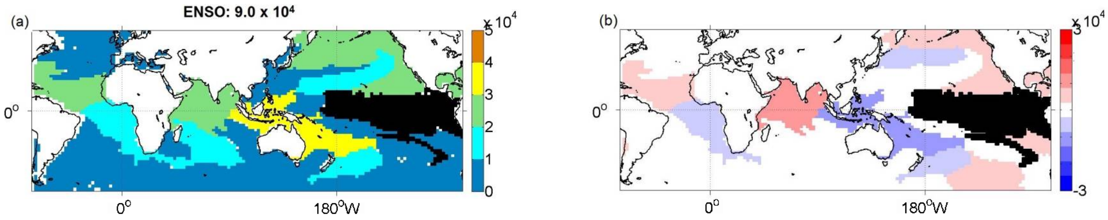


Figure 1: *Strength (a) and ENSO-related link (b) maps for the networks calculated using the HadISST during the period 1956-2005 in boreal summer (JJA). The strength of the ENSO-related area exceeds the colorscale and is saturated. Its value is indicated at the top of the stregnth panel.*

It was found that while SST data sets are characterized by comparable distances and ARI, precipitation reanalyses are critically different.

Application to CMIP5 Models

Secondly we analyzed a suite of 14 CMIP5 models (for clarity only 5 are shown in Figure 2 below; the complete analysis is available at Fountalis et al., 2015). The CMIP5 models have been validated against reanalyses over the second half of the 20th century, and compared for their projected responses under high GHG concentrations.

We focused on global quantities, and analyzed 50 years time intervals; the dominant mode of variability at those time and spatial scales is ENSO, which induces the most severe global impacts in surface temperatures and precipitation, among other variables. Despite decades of research, ENSO sensitivity to changes in GHG concentrations remains undetermined in the last generation of climate models. The results of our analysis can be summarized as follows:

- Within the CMIP5 inventory, several models reproduce closely the observed SST network over the historical period (1956–2005), providing an accurate representation of major modes of climate variability and their links, despite biases in the climatologies. The spread in ARI and D between SST networks from ensemble members of the same model is broadly consistent with the spread between different observational datasets or renalyses. Precipitation networks, unsurprisingly, indicate that spatial likeness and strength are still challenging for modelers. However, the limited agreement between reanalyses products, and the evaluation of the noise-to-signal ratio suggest that the spatial and temporal intermittency of precipitation intrinsically limits the reproducibility of its topology. Together those outcomes suggest that CGCMs cannot yet capture the observed natural variability of rainfall.

Models characterized by large D and small ARI in their SST fields, are also inaccurate in the representation of precipitation, but model performing the closest to the reanalysis in each of those fields differs.

- Changes in the network properties between the second half of the twentieth and twenty-first centuries are generally modest and contained within the spread between different observational proxies in the historical period, despite substantial trends. This is especially true for the models that reproduce accurately the recent past. For those models uncertainties are greater in the projected trends than in the response of their modes of variability (Fig. 2).

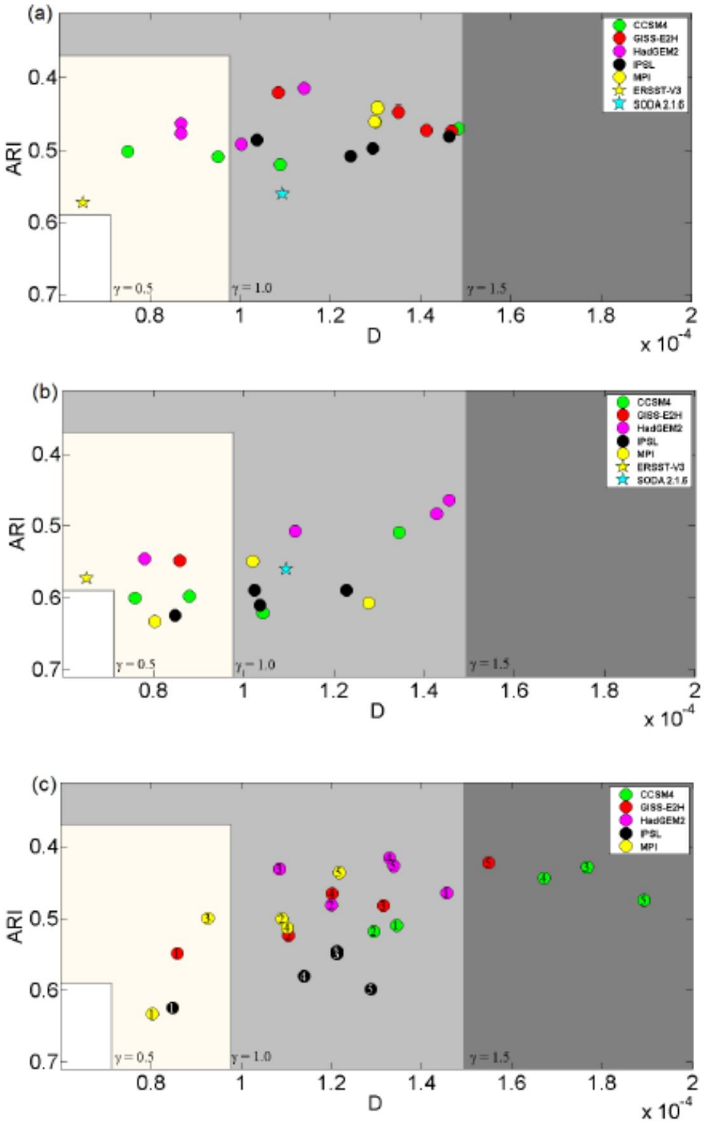


Figure 2. Metric D versus ARI for networks constructed using JJA sea surface temperature fields in subset of 5 models participating to CMIP5 (a) during the period 1956–2005, for up to 4 ensemble members for each model, (b) during 2051-2100 for the same members, and (c) from 2051 to 2300 over five consecutive 50-year periods, from 1 to 5, for the only ensemble member extending past 2100. In the historical period networks are referenced to the HadISST and the metrics are also indicated for two other reanalysis products. In the projected simulation all networks are referenced to the corresponding integration over the historical period. In the middle panels the metrics for the reanalysis products are repeated for context. Three levels of noise-to-signal ratios γ are also indicated.

Differences are slightly more probable in strength than in the spatial distributions of areas. Changes in distance D greater than 30 % around the historical value signals the model tendency towards strengthening or weakening of major climate modes, and of ENSO, its variance, and its connectivity, even when limited to one ensemble member. Eight of the twelve models analyzed display substantially weaker tropical areas and connections in one or more members, implying a decrease in ENSO strength and in potential predictability at seasonal and longer scales in the future. Only two models, MPI and MIROC5, display a clear trend towards intensifying the strength of the ENSO area and its links.

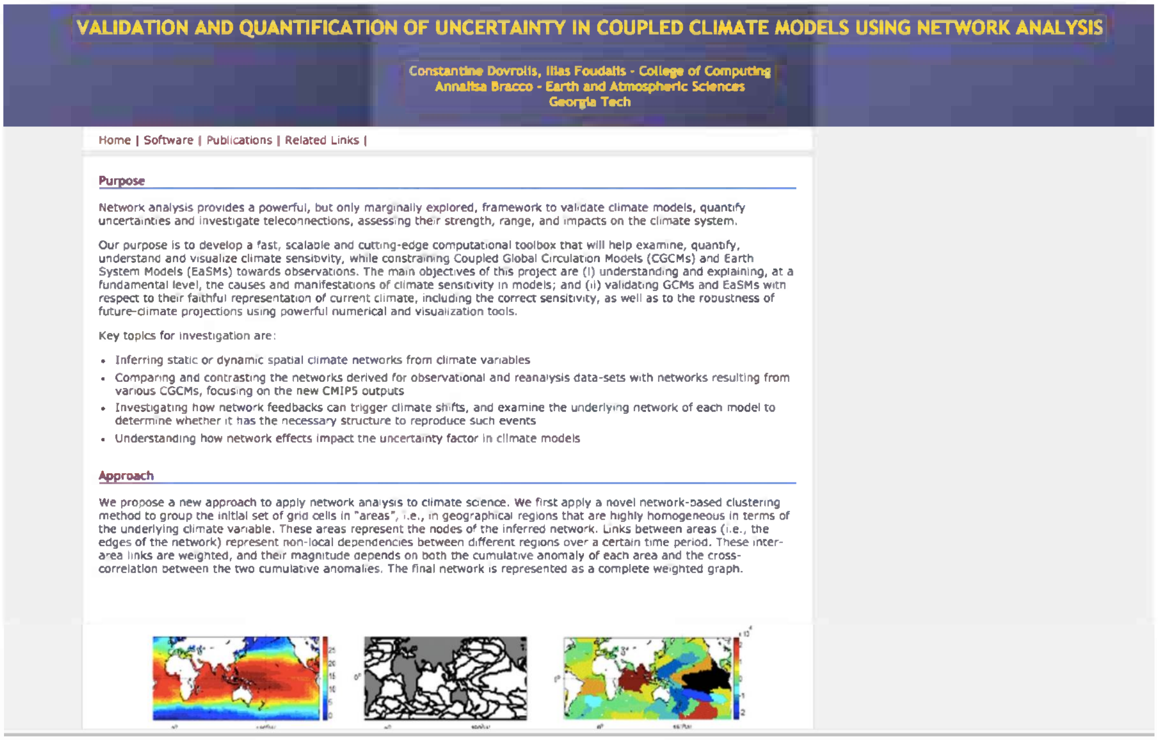
- After 2100, models forced by the concentration pathway of the scenario with the highest greenhouse gas concentrations in CMIP5 reveal discernible changes in the strength of all major areas. Five out of seven follow an irreversible trajectory towards reducing dramatically the strength and size of the ENSO node, and towards weakening all ENSO links over the 23rd century. This behavior is mirrored in precipitation to a lesser extent.

IPSL weakens as well, but partially recovers by 2300 in both SST and rainfall. MPI by the end of the integration has a virtually unaltered network in SST, while strength and links of the ENSO area increase substantially for precipitation.

Considering the global impacts of tropical teleconnections and the changes in temperature and precipitation associated with El Niño and La Niña events, we conclude that the uncertainty in the projected connectivity of the climate system after 2100 in many regions and for models performing well under current conditions exceeds the uncertainty associated with the equilibrium temperature change.

(E) Web-site

As part of our outreach activities and internal management of the project we have developed a project website (http://climatenets.org). This website is intended to be a live up to date reference point for the project activities and findings and allows for direct download of the software required to perform the network analysis proposed.



VALIDATION AND QUANTIFICATION OF UNCERTAINTY IN COUPLED CLIMATE MODELS USING NETWORK ANALYSIS

Constantine Dovrolis, Ilias Foudalis - College of Computing
Annalisa Bracco - Earth and Atmospheric Sciences
Georgia Tech

Purpose

Network analysis provides a powerful, but only marginally explored, framework to validate climate models, quantify uncertainties and investigate teleconnections, assessing their strength, range, and impacts on the climate system.

Our purpose is to develop a fast, scalable and cutting-edge computational toolbox that will help examine, quantify, understand and visualize climate sensitivity, while constraining Coupled Global Circulation Models (CGCMs) and Earth System Models (EaSMs) towards observations. The main objectives of this project are (i) understanding and explaining, at a fundamental level, the causes and manifestations of climate sensitivity in models; and (ii) validating GCMs and EaSMs with respect to their faithful representation of current climate, including the correct sensitivity, as well as to the robustness of future-climate projections using powerful numerical and visualization tools.

Key topics for investigation are:

- Inferring static or dynamic spatial climate networks from climate variables
- Comparing and contrasting the networks derived for observational and reanalysis data-sets with networks resulting from various CGCMs, focusing on the new CMIP5 outputs
- Investigating how network feedbacks can trigger climate shifts, and examine the underlying network of each model to determine whether it has the necessary structure to reproduce such events
- Understanding how network effects impact the uncertainty factor in climate models

Approach

We propose a new approach to apply network analysis to climate science. We first apply a novel network-based clustering method to group the initial set of grid cells in "areas", i.e., in geographical regions that are highly homogeneous in terms of the underlying climate variable. These areas represent the nodes of the inferred network. Links between areas (i.e., the edges of the network) represent non-local dependencies between different regions over a certain time period. These inter-area links are weighted, and their magnitude depends on both the cumulative anomaly of each area and the cross-correlation between the two cumulative anomalies. The final network is represented as a complete weighted graph.

(F) Publications resulted from the Grant

- Published:

1. Fountalis, I., Bracco, A., Dovrolis, C. (2015) ENSO in CMIP5 simulations: network connectivity from the recent past to the twenty-third century, Climate Dynamics, 45, 511-538, doi:10.1007/s00382-014-2412-1
2. Foudalis I., Bracco A. Dovrolis C. (2014) Spatio-temporal network analysis for studying climate patterns Climate Dynamics, 42, 879-899, doi:10.1007/s00382-013-1729-5

- In Press

1. Bracco, A., Archibald, R. K., Dovrolis, C., Fountalis, I., Luo, H., Neelin J.D. (2015) The parameter optimization problem in state-of-the-art climate models and network analysis for systematic data mining in model intercomparison projects. Book Chapter in CISM Courses and Lectures: The Fluid Dynamics of Climate, Springer Ed.

- In Preparation (Draft available)

1. Fountalis, I., Dovrolis, C., Bracco A., Dilkina B., Keilholz S., (In Prep.) A network-based method for the analysis of spatio-temporal data. Applications in climate and brain sciences. To be submitted to "The 20th Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2016" (Paper Submission Date October 2nd, 2015; http://pakdd2016.pakdd.org)

2. Fountalis, I., Bracco A., Venkatesan, H., Dovrolis, C., (In Prep.) The decadal modulation of ENSO statistics from a network prospective: Do CMIP5 models capture the observations? To be submitted to Geophysical Research Letters.

- Technical Report:

1. Fountalis I, Dovrolis C, Bracco A (2012) Efficient algorithms for the detection of homogeneous areas in spatial and weighted networks. Tech. Rep. College of Computing, Georgia Tech

(G) Presentation and Awards

1) 2012 AMS annual meeting (New Orleans, LA, January 2012), Theme Session 20: Data Mining, Prediction, and Predictability of the Climate System. Oral presentation, https://ams.confex.com/ams/92Annual/flvgateway.cgi/id/19860?recordingid=19860. Presenter: Ilias Foudalis. The presentation won the student presentation award for the session.
2) 3rd Workshop on Complex Networks (Complenets 2012) (Melbourne, FL, March 2012). Oral presentation. Presenter Ilias Foudalis.
3) Invited Seminar, Oak Ridge National Laboratory (ORNL) (Oak Ridge, TN, May 2012). Presenter Annalisa Bracco
4) AGU Fall Meeting (San Francisco, CA, December 2012), Poster presentation. Contributed. Presenter Annalisa Bracco
5) Third International Workshop on Climate Informatics, National Center for Atmospheric Research (Boulder, CO, September 2013). Invited Oral Presentation. Presenter Ilias Fountalis.
6) Invited Seminar, Un. California Los Angeles,. Dept. of Atmospheric and Oceanic Sciences (Los Angeles, CA, December 2014). Presenter Annalisa Bracco
7) Complex Network and Climate Variability Conference (Vienna, Austria, April 2015). Invited Oral Presentation. Presenter Annalisa Bracco