5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

TELECONNECTION SIGNALS EFFECT ON TERRESTRIAL PRECIPITATION: BIG DATA ANALYTICS VS. WAVELET ANALYSIS

Yahui Di[1], Wei Ding[1], Sanaz Imen[2], Ni-Bin Chang[2]

Abstract— **The main purpose of this study is to determine the association rules between hydro-climatic variables and the atmospheric / oceanic variables separated by large distances, which are known as the phenomenon of hydro-climatic teleconnection. In order to discover physically meaningful patterns from big climate databases, we aim at developing efficient data-driven approaches with the aid of machine learning, signal processing, and domain knowledge for constrained search. The big data analytics tool with the streaming feature selection in machine learning extracts hydro-climatic variables from large temporal and spatial feature space and formulates the global search for teleconnection signals effect on terrestrial precipitation. The wavelet analysis retrieves the scale-averaged wavelet power to signify the teleconnection signals via a pixel-wise linear lagged correlation analysis. Preliminary comparisons between streaming feature selection in machine learning and wavelet analysis were made possible to pin down some known teleconnection patterns in this interdisciplinary study.**

I. INTRODUCTION

The main purpose of this study is to determine the association rules between hydro-climatic variables and the atmospheric / oceanic variables separated by large distances, which are known as the phenomenon of hydro-climatic teleconnection. In order to discover physically meaningful patterns from big climate data, we aim at developing efficient data-driven approaches with the aid of machine learning, signal processing, and domain knowledge for constrained search.

Big Data Analytics. From the machine learning perspective, we propose to employ the big data analytics that extracts hydro-climatic variables from large global temporal and spatial feature space and formulates the problem of the teleconnection signal effect on terrestrial precipitation variability as feature selection in machine learning. Because of the air mobility and the closed climate system [1] [2], we took the spatial and temporal influences into consideration when constructing the feature set. As suggested by the literature [2], Sea Surface Temperature (SST) has lag effect via signal propagation to affect the terrestrial precipitation variability, which is about 1~12 months. The global SST with a time lag of 1 to 12 months (one year) is thus adopted as the features for seasonal precipitation analysis. In order to establish the possible association rules between SSTs forcing and precipitation responses, seasonal time series were calculated by computing SST averages over every three months on a seasonal scale, namely March-April-May for spring, June-July-August for summer, September-October-November for fall, and December-January-February for winter. Therefore, 13 different SSTs time series were computed with time lags from 0 to 12 months. For instance, to find the lagged correlation between precipitation of March-April-May, the associated SST time series are averaged for three month periods starting with March-April-May (0 month lag), February-March-April (1 month lag), January-February-March (2 months lag), ..., March-April-May (12 months lag). In doing so, our first scenario ended up with building a group of 374,400 features in each possible month (28,800 locations time 13 months) if we choose 1.5×1.5 degree resolution on the Northern and Southern hemispheres of the Earth. This is a big-data problem over extremely high feature dimension, whereas many of the constructed features from this huge spatio-temporal space could be irrelevant and redundant. It calls for an efficient method to select the features which are most-relevant to the precipitation event for a given location.

Wavelet Analysis. Our domain knowledge also led us to use wavelet analysis over the spectral dimension to search for nonlinear and non-stationary time series signals

Corresponding author: Y. Di, University of Massachusetts Boston, Boston, MA yahui.di@gmail.com [1]Department of Computer Science, University of Massachusetts Boston, Boston, MA [2]Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL

embedded in the climate teleconnections from which those index regions may be more easily retrieved by a subsequent pixel-wise linear lagged correction analysis [1][2]. In contrast, the Big Data Analystics method that Ding et al. used is an efficient streaming feature selection method [3] to identify strongly relevant non-redundant features from extremely high feature space. Comparison of the results from the two methods to discern the meaningful teleconnection patterns from big climate data is the focus.

II. METHODS

The goal of Fast-Online Streaming Features Selection (Fast-OSFS) is therefore designed to find strongly relevant and non-redundant features from an extremely high dimensional dataset [3]. The method builds a local Bayesian network and efficiently selects strongly relevant features using Markov blankets. We processed the large volume of features sequentially and evaluated one feature at a time—a feature is evaluated upon its arrival and is decided whether it can be added into the current pool of selected features. To avoid the expensive calculation of a full Bayesian network, our algorithm discovers Markov blankets for each feature seen so far using the direct causes and direct effects that produces the skeleton of a Bayesian network.

The streaming feature selection method that is developed by the UMASS-Boston team is an efficient data-driven approach. But the method neglects domain constrains, thus it is possible that physical meaningful patterns could be removed during the Online Redundancy Analysis. The UCF team used integrated wavelet spectral analysis and Pearson correlation-based regression analysis to search for nonlinear and non-stationary signals of climate teleconnections associated with terrestrial sites [2]. The basic analytical framework is shown as Figure 1. It operates in two phases. In phase I, it identifies the possible index regions. In phase II, it screens and ranks the identified index regions.

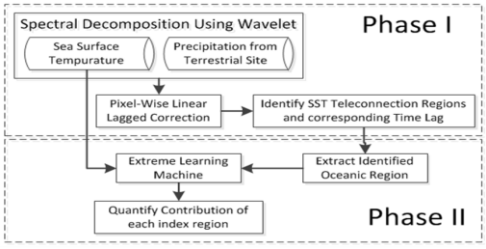

Figure 1 The analytical framework of Wavelet Analysis

III. PRELIMINARY RESULTS

We chose the Adirondack, NY as the study area. The precipitation data in Adirondack is a full data product of the Global Precipitation Climatology Center (GPCC-V6), and the global SST is the ERA-Interim reanalysis product. The preliminary holistic comparison among teleconnections (NOAA ocean indices) and two teams' findings are shown in Figure 2. In the four maps, the found index regions by two teams are somewhat overlapped, and most of them are located in the area of NOAA ocean indices. However, some of the found index regions on the maps are quite different, like those in the fall. We will keep on studying what caused the selection of these different regions and which regions have higher contribution to the precipitation variability at Adirondack site.

IV. CONCLUSION

In this paper, we present our preliminary results with respect to the index regions based on the streaming features for big data and wavelet analysis for teleconnection signal analysis. Our next research work will be the integration between streaming feature selection and wavelet analysis to be more efficiently identify physically meaningful teleconnection patterns from big climate data.

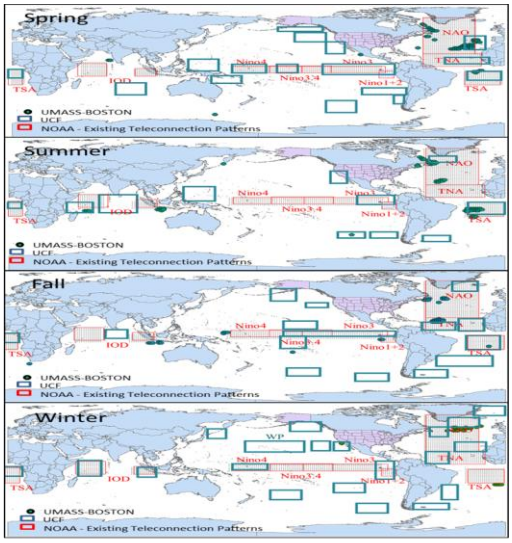

Figure 1 The comparison among NOAA-Existing Teleconnection Patterns and two different method's findings. The "UMASS-BOSTON" represents then indices founded by using streaming feature selection method in Ding's team. The "UCF" means the indices founded by using the wavelet analysis in Chang's team.

REFERENCES

[1] N. B. Chang, M. V. Vasquez, C.F. Chen, S. Imen, and L. Mullon. "Global nonlinear and nonstationary climate change effects on regional precipitation and forest phenology in Panama, Central America." Hydrological Processes 29, no. 3 (2015): 339-355.

[2] L. Mullon, N. B. Chang, Y. J. Yang, and J. Weiss. "Integrated remote sensing and wavelet analyses for screening short-term teleconnection patterns in northeast America." Journal of Hydrology 499 (2013): 247-264.

[3] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu. "Online feature selection with streaming features." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35, no. 5 (2013): 1178-1192