POLITECNICO DI MILANO

Scuola di Ingegneria Civile, Ambientale e Territoriale

Master of Science in

Environmental and Land Planning Engineering

Assessing the effects of Sea Surface Temperature Anomalies in the hydrology of the Red River Basin, Vietnam

Supervisor:

PROF. ANDREA CASTELLETTI

Co-Supervisor:

PROF. STEFANO GALELLI

Master Graduation Thesis by:

YLENIA CASALI
Student Id n. 799380

Academic Year 2013-2014

POLITECNICO DI MILANO

Scuola di Ingegneria Civile, Ambientale e Territoriale

Corso di Laurea Magistrale in
Ingegneria per l'Ambiente e il Territorio

Valutare gli effetti delle anomalie di Sea Surface Temperature nell'idrologia del bacino del Fiume Rosso, Vietnam

Relatore:

PROF. ANDREA CASTELLETTI

Correlatore:

PROF. STEFANO GALELLI

Tesi di Laurea Magistrale di:

YLENIA CASALI
Matricola n. 799380

Anno Accademico 2013-2014

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CCA	Canonical Correlation Analysis
EOF	Empirical Orthogonal Functions
ENSO	El Niño Southern Oscillation
ERSST	Extended Reconstructed Sea Surface Temperature
ET	Extremely randomized Trees
ICOADS	International Comprehensive Atmosphere-Ocean Data Set
IIS	Iterative Input variable Selection
IOD	Indian Ocean Dipole
IR	Input Ranking
IVS	Input Variable Selection
MAI	Most Anomalous Indicator
MB	Model Building
MEI	Multivariate ENSO Index
MISO	Multi Input-Single Output
OI	Optimum Interpolation
PA	Parallel Analysis
PCA	Principal Component Analysis
PDO	Pacific Decadal Oscillation
SISO	Single Input-Single Output
SO	Southern Oscillation
SOI	Southern Oscillation Index
SST	Sea Surface Temperature
SSTA	Sea Surface Temperature Anomalies

1 INTRODUCTION

Over the past several decades, hydrologists and climatologists have developed relationships between large scale oceanic–atmospheric variability and climate [*Soukup et al.*, 2009]. Studying the relationship between ocean-atmospheric interaction and hydro-meteorological processes is helpful in hydrologic and meteorological forecasting and water resources management. One way of studying these relationships is to use the ocean's Sea Surface Temperature (SST) [*Meidani and Araghinejad*, 2013], since SST affects hydro-meteorological processes by teleconnections. It is known that some patterns of SST affect hydro-meteorological variables by a sea-land-atmosphere interaction known as teleconnection. Moreover low frequency climatic fluctuations, related to SST variations, are known to be a major factor causing extreme hydrological events and significant variations in water resources at global and regional scales [*Kahya and Dracup*, 1993]. Therefore SST variability can provide important predictive information about hydrologic variability in regions around the world [*Soukup et al.*, 2009]. Several works can be found in the literature investigating the relationship between SST and hydro-meteorological processes throughout the world [*Barnston and Smith*, 1996]. These works try to assess this relationship by employing SST indexes from ocean areas that have a well-studied role in regulating climate and effects at global and regional scales. SST indexes are then used to make prediction models in order to improve the anticipation capability of hydro-meteorological processes. Actually a purpose of water resources management is to extend the lead time in order to achieve medium-to-long range forecasts. If reliable medium-to-long range forecasts of streamflow were available, water management policies accounting for flood control, power generation and irrigation water supply could be developed. Knowing the resource available months in advance, measures for adaptation and resilience in water uses can be arranged leading to an overall improvement in water resources management [*Savage*, 2013].

Significant researches have focused on identifying atmospheric–oceanic climatic phenomenon of the El Niño Southern Oscillation (ENSO) [*Soukup et al.*, 2009]. Most likely the best known teleconnection phenomenon, ENSO is a large-scale coupled ocean–atmosphere phenomenon occurring in the Tropical Pacific Ocean that has been linked to climate anomalies throughout the world [*Chiew and McMahon*, 2002]. ENSO effects are studied mostly in certain regions that are well known for their sensitivity to the ENSO phenomenon, for example in the coastal regions of Northern Peru and Southern Ecuador. In general, several researchers have shown a significant relationship between ENSO events and the rainfall of Pacific Rim countries and of the tropical belt regions [*Gutierrez and Dracup*, 2001 and *Amarasekera et al.*, 1997]. ENSO effects on hydrometeorological processes are studied using standard indexes, such as averaged SST over particular regions of the Pacfic Ocean (e.g. Niño3, Niño4, Niño3.4, Niño1+2) or other indexes related to other variables in that region (e.g. the Southern Oscillation Index (SOI) and the Multivariate ENSO Index (MEI)). Often, however, the standard indices of these phenomena are not good predictors of hydroclimate in every basin of the Pacific Rim, for example in the western United States, even though ENSO phenomenon impact that areas hydroclimate. Because the canonical patterns of this climate phenomenon refer to specific regions in the ocean and slight shifts in the patterns can result in decreased correlation values between the indices and basin hydroclimate [*Grantz and Rajagopalan*, 2005]. Therefore, depending on what region is being examined, it is possible that other indicators may be more appropriate. Alternative indices of ENSO phenomenon are studied to find a stronger ENSO-hydrology correlations.

New indexes are found by averaging SST over the areas of highest correlation with the basin hydrology (e.g. streamflow). These areas are determined by visual inspection of the correlation maps or by test of significance. For example, good results are obtained in western United States, averaging Pacific SST [*Grantz and Rajagopalan*, 2005]. Other researches follow this method in ocean area that not concern only ENSO phenomenon. For example Phillips et al. have found indexes in North Atlantic SST that can forecast precipitation in Iceland [*Phillips and Thorpe*, 2006]. Researches of Tarakanov and Borisova developed recently Most Anomalous Indicator (MAI). In fact recent advances in satellite remote sensing techniques are making possible to identify new oceanic regions that could be used as ENSO SST indicators. MAI can serve as a convenient indicator of extreme oceanographic conditions and can be also used as a predictor of global events [*Tarakanov and Borisova*, 2013]. Further researches use statistical tools for identifying coupled relationships between SST and hydrometeorological variables. For example, Roswiarti et al. use Empirical Orthogonal Functions (EOF) and Canonical Correlation Analysis (CCA) to extract the impact to El Niño on North Carolina precipitation. Results provide a confidence in the applicability of EOF and CCA analysis for understanding El Niño-like climatic events under regional or local perspectives [*Roswintiarti, Devdutta, and Raman*, 1998].

Moreover it is important to note that basin hydroclimate could be not strongly related to specifically ENSO phenomenon more than other teleconnection phenomena. Therefore, alternative indices can be found assessing the relationship between hydroclimate variables in a specific basin with oceanic areas that not concerns only ENSO area. It is therefore clear that a great deal of uncertainty and significant subjectivity exist in selecting not only which index to use but also which oceanic areas are the most appropriate to study.

1.1 OBJECTIVES

The purpose of this thesis is to assess the effects of SST of Indian and Pacific Ocean on the hydro-meteorological processes over the Red River Basin, Vietnam. This work is intentionally focused to cover the impacts of Indian and Pacific Ocean on the hydrology of the Red River Basin, because unambiguous research evidence showing these connections are not found. Although the ENSO–hydrology relationship is found to exist in Vietnam, there are still several open research questions to be answered [*Räsänen and Kummu*, 2013].

The main goals of the present thesis are to assess the impact of SST on the Red River Basin, means by to evaluate of new indicators of SST that compress the main informations and behaviour of the SST in the oceans, and to develop streamflow prediction models incorporating the SST indicators found in the previous analysis.

In order to achieve these goals, this work presents a framework to identify large-scale climate indicators of Sea Surface Temperature Anomalies (SSTA) related to hydro-meteorological variables of the basin and to incorporate this indicators into forecasts of streamflow. The framework follows four main steps. Each step achieves a particular objective, in order: the first step assess the relationship between SST and hydro-meteorological variables, the second step create indicators of SSTA, the third step assess the relationship between the new sets of SSTA indicators and hydro-meteorological variables, the fourth step compute prediction model by using the information obtained in the previous steps. The findings of every step are employed in the following step. Each step is described above:

1. Correlation analysis is adopted to prove the relationship between Indian and Pacific SST and hydro-meteorological variables of the Red River Basin. Squared grid SST data sets are used. Each grid point of SST is correlated with streamflow and rainfall series of the basin. Correlation maps obtained show the areas strongly correlated.

2. Empirical Orthogonal Functions (EOF) of SSTA is calculated in order to create new sets of variables that could be used as indicator of SSTA. The advantage of using EOF is to compress the information of the input variables, such as SSTA, in order to obtain new sets of independent variables. It is possible to select few independent variables in order to explain most of the total variance of the process.

3. Canonical Correlation Analysis (CCA) between selected EOFs of SSTA (predictors) and EOFs of rainfall/streamflow (predictands) is computed. CCA is a powerful statistical tool for identifying coupled relationships between predictors and predictands. CCA correlates linear combinations of a set of predictors that maximize relationships, in a least-square sense, to linear combinations of predictands [*Roswintiarti, Devdutta, and Raman*, 1998]. Therefore CCA is used to assess the relationship between the indicators of SSTA and the hydro-meteorological variables.

4. Prediction models using the indicators of SSTA are analyzed. Two different methods are used: CCA and Input Variable Selection (IVS). CCA and IVS are computed between the indicators of SSTA and the streamflow anomalies at each station of the basin. In this step CCA is used in a predictive way. It's possible to use the linear combinations obtained from the method to make predictions. In general IVS is used in order to find the most relevant climatic forcings of at site streamflow variability and to derive a predictive model based on the inputs selected. Therefore IVS is used first to evaluate if SST indicators are predictors of streamflow anomalies and second to build prediction models of the selected SST indicators.

In summary, the new contribution of this thesis is the development of a framework in which indicators of SST are assessed, in order to improve prediction methods for teleconnection induced hydrological anomalies.

The present thesis is structured into two distinct parts. In the first part, the theoretical background is provided.

- Chapter 1 contains an overview on the role of SST on hydro-meteorological processes and teleconnection phenomena. Additionally some SST data sets are described.

- Chapter 2 describes the methods and tools adopted in this work.

In the second part, case study and results are shown.

- In Chapter 4 the Red River Basin and its hydro-meterological variables are described.

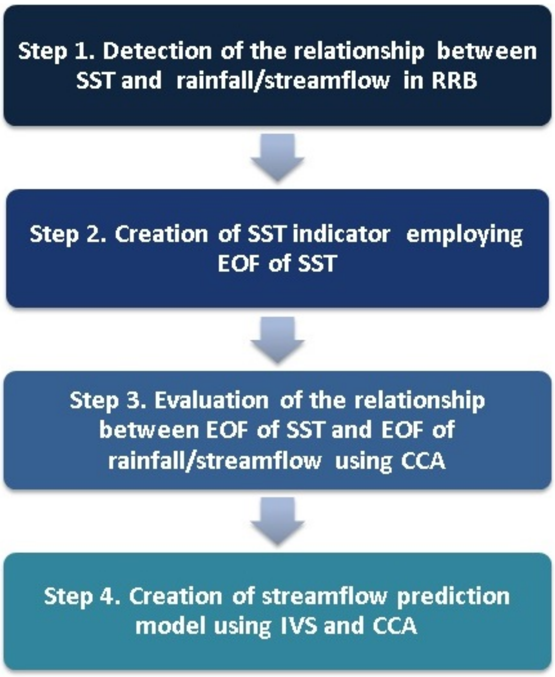
Step 1. Detection of the relationship between SST and rainfall/streamflow in RRB
Step 2. Creation of SST indicator employing EOF of SST
Step 3. Evaluation of the relationship between EOF of SST and EOF of rainfall/streamflow using CCA
Step 4. Creation of streamflow prediction model using IVS and CCA

Figure 1.1: Flowchart of the analysis

- Chapter 5 contains results and discussion of the impact of SSTA on hydro-meteorological processes.
- Chapter 6 contains results and discussion of the prediction models.

Conclusions and fruther enhancements are drawn in Chapter 7.

Part I

THEORY

2 IMPACT OF SST ON CLIMATE AND HYDRO-METEOROLOGICAL PROCESSES

The purpose of this chapter is to describe the SST variable and its effect on climate and hydro-meterorological processes.

2.1 SST AND CLIMATE

In recent decades, studies on sea-land-atmosphere interaction have a crucial role in the understanding of global climate. In the dynamic of hydrological cycle, oceans play an important role owing in part to their large heat-storage capacity. Physically the ocean' s thermal inertia is linked to the atmosphere via turbulent and radiative energy exchange at the sea surface. These energy fluxes, in turn, depend on a single oceanic quantity, the sea surface temperature, as well as several atmospheric parameters including wind speed, air temperature, humidity and cloudiness. Since 1960s many studies have confirmed that SSTs play a key role in regulating climate and its variability. In particular, slow variations in SST provide a source of potential predictability for climate fluctuations on timescales of seasons and longer. SST behavior differs fundamentally from atmospheric variables (e.g. pressure), which often show marked variability at the daily timescale and are strongly linked to concurrent but not future rainfall variability [*Deser et al.*, 2010].

It is known that some pattern of SST are related to atmospheric circulation variability by teleconnections. A phenomena of teleconnection refers to recurring and persistent patterns of climate variables located in two different places, during which one variable affects the other one at large distance. Due to its relationship with teleconnections phenomena, SST affects the hydro-meteorological processes at the basin scale. Teleconnections phenomena are widespread in different geographical areas and they can sometimes be prominent for several consecutive years, thus reflecting an important part of both the interannual and interdecadal variability of the atmospheric circulation. Famous examples of climate indices linked to teleconnection patterns of atmospheric circulation variability include the Pacific Decadal Oscillation (PDO), the Indian Ocean Dipole (IOD) and the El Niño Southern Oscillation (ENSO). A brief description for each phenomenon is provided below. ENSO is addressed in a dedicated section.

2.1.1 *PDO and IOD*

The Pacific Decadal Oscillation (PDO) occurs in the North Pacific Ocean, where SST changes between warm (positive values) and cool (negative values) phases occur every 20 to 30 years. When SSTs are anomalously cool in the interior North Pacific and warm along the North American Pacific coast, and when sea level pressures are below average over the North Pacific, the PDO has a positive value. When the climate anomaly patterns are reversed, with warm SST anomalies in the interior and cool SST anomalies along the North American coast, or above average sea level pressures over the North Pacific, the PDO has a negative value. PDO affects specially the North American climate. Warm phases of the PDO are correlated with above average winter and spring time temperatures in northwestern North America, below average temperatures in the southeastern United States, above average winter and spring rainfall in the southern United States and northern Mexico, and below average precipitation in the interior Pacific Northwest and Great Lakes regions. Cool phases are correlated with the reverse climate anomaly patterns over North America. The PDO-related temperature and precipitation patterns are also strongly expressed in regional snow pack and stream flow anomalies, especially in western North America.

The Indian Ocean Dipole (IOD) is centered in the equatorial Indian Ocean and it affects mainly the climate of countries that surround the Indian Ocean basin, Australia included. It is commonly measured by an index that is the difference between SST anomalies in the western (50°E to 70°E and 10°S to 10°N) and eastern (90°E to 110°E and 10°S to 0°S) equatorial Indian Ocean. The index is called the Dipole Mode Index (DMI). A positive IOD period is characterised by cooler than normal water temperatures in the tropical eastern Indian Ocean and warmer than normal water temperatures in the tropical western Indian Ocean. Conversely, a negative IOD period is characterised by warmer than normal water temperatures in the tropical eastern Indian Ocean and cooler than normal water temperatures in the tropical western Indian Ocean. Some effects of IOD are observed over Australia, where a positive IOD SST pattern is associated with a decrease in rainfall over parts of central and southern Australia. Otherwise a negative IOD SST pattern is associated with an increase in rainfall over parts of southern Australia.

2.1.2 *ENSO*

The El Niño Southern Oscillation (ENSO) is centered in the equatorial Pacific Ocean. It consists of two phases, the warm El Niño phase and the cold La Niña phase, that are connected to the atmosphere through a seesaw atmospheric pressure fluctuation in the South Pacific called the Southern Oscillation (SO) [*Shrestha and Kostaschuk*, 2005]. Typically, ENSO events occur at irregular intervals, with a characteristic return frequency of 2-7 years (which makes ENSO a quasi-periodic phenomenon) and usually persist for 1–2 years. No two events are completely alike: they evolve according to a consistent pattern, but they differ in timing, intensity, extent, and duration [*Kahya and Dracup*, 1993]. ENSO events are related to inter-annual variations in precipitations and streamflow in several regions of the world. Therefore the ability to predict flow patterns in rivers and precipitation forecasting in a certain region will be highly enhanced if a strong relationship between ENSO-climate variables is quantified [*Amarasekera et al.*, 1997]. Seeing the potential offered by knowledge of ENSO events, many researchers assessed the relationship among ENSO event and hydrometeorological processes all over the world [*Barnston and Smith*, 1996]. Several indexes to monitor ENSO condition are available, and the most widely used are the SOI and the SST based indexes, calculated over different areas of the tropical Pacific Ocean known as Niño 1+2, Niño 3, Niño 3.4 and Niño4 [*Kiem and Franks*, 2001].

SST-based indexes have been derived using different areas of the equatorial Pacific Ocean (see figure 2.1):

- Niño1+2: 80°W – 90°W, 10°S – Equator;
- Niño3: 90°W – 150°W, 5°S – 5°N;
- Niño3.4: 120°W – 170°W, 5°S – 5°N;
- Niño4: 150°W – 180°W, 5°S – 5°N.

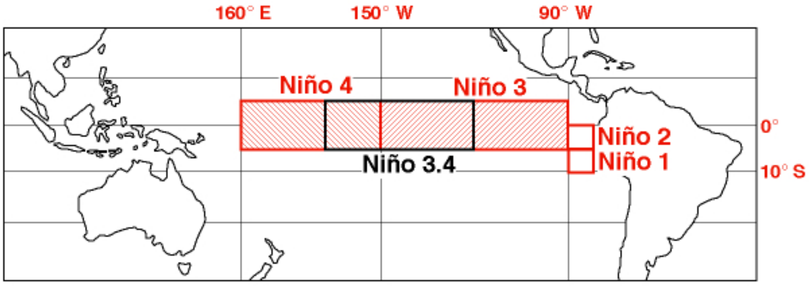

Figure 2.1: Equatorial Pacific SST regions. Source: National Center for Atmospheric Research (`http://www.ucar.edu/communications/newsreleases/1998/ninatip.html`).

The SSTs characterizing the identified regions are averaged for each corresponding ENSO index [*Chandimala and Zubair*, 2007]; for example, the Niño3 index is defined as the seasonal SST averaged over the

Niño3 region, which is located in the central/eastern Pacific Ocean between 5°S and 5°N in latitude and 90°W and 150°W in longitude.

More recently, other ENSO indexes have come to the attention of researchers, such as the Multivariate ENSO Index (MEI) and Most Anomalous Indicator (MAI). The MEI is based on six observed variables over the tropical Pacific Ocean. These six variables are Sea Level Pressure, Sea Surface Temperature, zonal and meridional components of the surface wind, Surface Air Temperature and total cloudiness fraction of the sky. According to several researchers, being derived from multiple climate variables, the MEI could potentially represent a more integrated measure of persistent anomalies and, therefore, reflect the nature of the coupled ocean-atmosphere system better than either SOI or SST based indexes. This is because integrating more information than the SOI and SST-based indexes, which are each based on a single variable (pressure and temperature, respectively), the MEI would result less vulnerable to non-ENSO related variability [*Kiem and Franks*, 2001]. Tarakanov and Borisova developed the MAI index that is defined as a geographical point where the value of SST is greater than in any other point of the area. Such singular point of the MAI can serve as a convenient indicator of extreme oceanographic conditions and can also be used as a predictor of global events [*Tarakanov and Borisova*, 2013].

2.2 SST AND HYDROMETEOROLOGICAL PROCESSES

Thanks to the effects of teleconnections, studies have been developed during the last decades focusing on the relationship among SSTs and meteorological variables of a geographical area in order to forecast hydro-meteorological processes. One easy way of measuring these relationships is to use SST indexes, alternatively a selection of a shorter portion of SST over oceans is used in order to evaluate which particular ocean area affects more the meteorological variables. It is a common approach to assess the influence of a particular ocean portion on a regional area by dividing the ocean in a grid square of SST values in order to study the influence of each square on the hydro-meteorological variables (e.g. rainfall, streamflow) recorded on that area in a specific lagged time. Instead of using SST, many studies use SSTA that are definied as the SST removed from its long-term average. It is possible to calculate SSTA averaging the SST at each month and removing the montly average from each month of SST samples. SSTA is useful because it is able to assess the influence of the sea surface temperature no affected by seasonal variability.

SSTs can be used to provide an indication of whether the following month or season is likely to be wetter or drier than average, but they cannot be used to predict individual extreme events at the synoptic timescale, that is the scale of an atmospheric condition. This restricts the utility of rainfall forecasts using SST because planning authorities will often be more interested in short-lived rainfall extremes than time-averaged behavior over a season. However, the development of globally complete SST and sea-ice data sets allows climate scientists to study SST–weather linkages with greater confidence than their predecessors, who were hindered by data sets that were geographically sparse, of limited record length and homogeneity [*Phillips and Thorpe*, 2006].

Many studies in literature focused particularly on the influence of tropical Pacific SST, such as the area where ENSO is the main pattern, on meteorological variables, because ENSO is an event of great interest in the climate field and, further, its strong influence on climate variables all over the world is well documented [*Chiew and McMahon*, 2002].

The work of Barnston and Smith (1996) attempts to use SST to examine relationships between SST and precipitation an temperatures over most of the world. They use Canonical correlation analysis (CCA) in order to provide a physical interpretation of the relationships for the 1950-1992 period. Specifically a sequence of four consecutive 3-months periods of global SST anomalies is related to temperature and precipitation anomalies during 3-month periods ranging from zero to four seasons later. Results show that specification and prediction are relatively skillful, first, in areas affected by the ENSO. These include the tropical Pacific islands and most of the tropical Indian Ocean land regions throughout most of the year. They include also portions of North and South America, Africa and Australia for specific seasons for both temperature and precipitation. In South America precipitation, it was noted that the Atlantic SST can be at least as important as Pacific SST. Results show that additional predictive skills are observed also from other processes than ENSO. Trends in non-ENSO sources appear to control the continental climate as strongly as ENSO both in regions directly affected by ENSO and those are minimally influenced [*Barnston and Smith*, 1996].

Many studies focus on Europe and often assess the relationships between Atlantic Ocean and different regional area around Europe. Lorenzo et al. (2010) looks at the relationship between monthly North Atlantic Sea Surface Temperatures Anomalies (SSTAs) and regional index of rainfall in northwest Iberian Peninsula during the period 1951-2006, similarly Phillips Thorpe (2006) assesses the relationship between gridded (5°× 5°) monthly North Atlantic (10–70°N, 80°W–20°E) SSTAs and concurrent, one-monthly and two-monthly lagged rainfall totals for four coherent Icelandic precipitation regions over the period 1961–2002. Both studies evaluate the strength of the relationships by correlation coefficient for any given grid square of SSTA with the meteorological variable on lands.

Roswiarti et al. (1998), studying the linear teleconnections of El Niño events and precipitation over North Carolina, found an highly correlation among monthly tropical Pacific SSTA and precipitations. In particular the results show that El Niño-related precipitation anomalies along the North Carolina coast were positive from November to May and negative between June and October consistent with large-scale studies.

Using winter monthly grid of SST in order to predict East Asian summer monsoon (EASM), Li and Zeng (2008) found that the forecast skill of EASM rainfall, based on CCA prediction model, is quite low although the tropical Pacific SST in winter has been generally recognized as the most prominent previous signals from the sea.

An example of a study in Arabic peninsula is taken from Meidani (2013), who investigates the rainfall and streamflow over southwestern Iran from SST of Mediterranean Sea. The aim is to evaluate the correlation of SST with the hydrological response of the study area through singular value decomposition (SVD). The analysis show that Mediterranean SSTs have a clear significant impact on wet season (February to May) average streamflow over the south west of Iran.

2.3 SST DATA

Nowadays some known organizations arrange open climate variables on the web; NOAA (National Oceanic and Atmospheric Administration) collects a widespread and updated archives of SST data set. Other organizations which provide analysis of SST data are the Australia Government Bureau of Meteorology and the Met Office Hadley Centre. In this section are considered those data sets that are available in a regular grid over the world oceans on NOAA's website. The SST data sets discussed below are reported in tab.2.1.

2.3.1 *ICOADS*

The International Comprehensive Atmosphere-Ocean Data Set (ICOADS) is an extensive and widely used digital collection of quality-controlled surface weather observations (including SST). The majority of the measurements come from ships of opportunity, supplemented in recent years by research vessels, moored environmental buoys, drifting buoys,

Name	Period of record	Spatial resolution	Comments
ICOADS	1/1800-present	2°x2°since 1800 1°x1°since 1860	In situ
Kaplan	1/1856-present	5°x5°	In situ/satellite
ERSST	1/1854-present	2°x2°	In situ/satellite
Optimum Interpolation	12/1981-present	1°x1°	Satellite estimates

Table 2.1: SST indexes, source NOAA (http://www.esrl.noaa.gov/psd/data/gridded/)

and near-surface measurements from hydrographic profiles. The data are monthly averaged and binned into 2°latitude by 2°longitude grid beginning in 1800 (1°by 1°beginning in 1960) and extending through 2007. Due to the uneven distribution of commercial shipping routes and changes in those routes over time, data coverage is poor in certain regions and periods (see figure 2.2). Specifically the North Atlantic, western South Atlantic, and northern Indian oceans contain the highest density of observations, with reasonable coverage back to approximately 1870. Data coverage is limited in the North Pacific before 1946 and in the tropics before 1960; the Southern Ocean remains poorly sampled throughout the record [*Deser et al.*, 2010].

The lack of spatial and temporal smoothing in the ICOADS, along with large uncertainties in individual monthly mean values due to inadequate sampling, makes it difficult to produce comprehensible maps for a specific month and year without additional processing of the data. Various empirical and statistically optimal procedures have been employed to improve upon the sampling uncertainties, temporal lack, and missing data in the ICOADS and related archives.

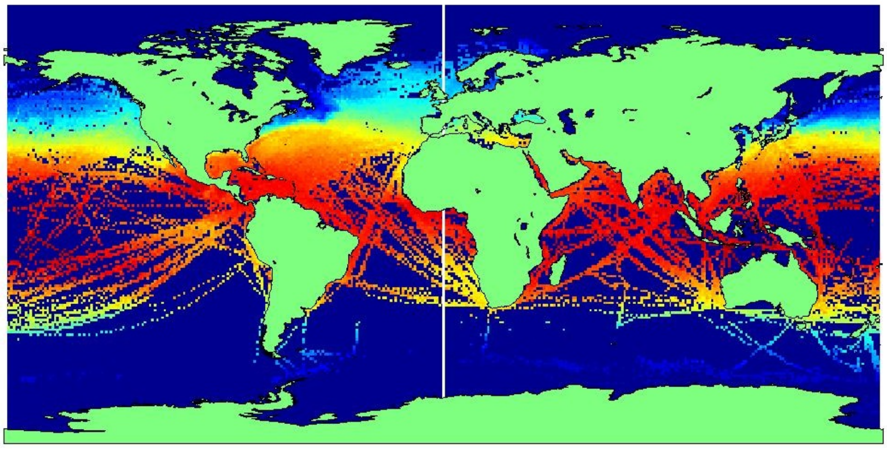

Figure 2.2: ICOADS SST data set at the beginning of 1970s

2.3.2 *Kaplan, ERSST, OI*

Globally complete monthly SST products (e.g. Kaplan, the Extended Reconstructed Sea Surface Temperature (ERSST) and the Optimum Interpolation (OI)) are useful in studies of global SST variability, although they remain constrained by the quality, quantity, and distribution of the original measurements. Thanks to the advent of remote sensing at the beginnings of the early 1980s, there is now full coverage over the world oceans at both high temporal (sample every few days) and spatial (1°latitude by 1°longitude) resolution. The satellite data are achieved by infrared sensors from the Advanced Very High Resolution Radiometer and by microwave measurements from the Advanced Microwave Scanning Radiometer; they are sometimes blended with conventional in situ data to account for biases.
One of the most recent SST product is the NOAA's OI, which has been collecting data since the beginning of 1980s. Also if there are not analysis over land, values are filled also over it by an interpolation in order to produce a complete grid. The OI analysis is produced weekly on a spatial resolution of one-degree grid and the monthly values are derived by averaging the weekly values over a month.
Kaplan dataset is derived from Met Office Hadley Centre SST data, which is an input to sophisticated statistical techniques applied to fill gaps. It has a high spatial resolution of five-degree grid and data are avaible only in a monthly sampling.

The ERSST data (see figure 2.3) constructed using the most recently available ICOADS SST data and improving statistical methods that allow stable reconstruction using sparse data. In this study, it is used the monthly ERSST data set for two mainly reason: the long temporal coverage (from 1854 to present) and the higher spatial resolution (2°latitude by 2°longitude). It has a more accurate spatial resolution compared with Kaplan (5°latitude by 5°longitude) and a longer temporal coverage compared with OI. These properties are essential in order to produce precise analysis over decades which start before 1980s.

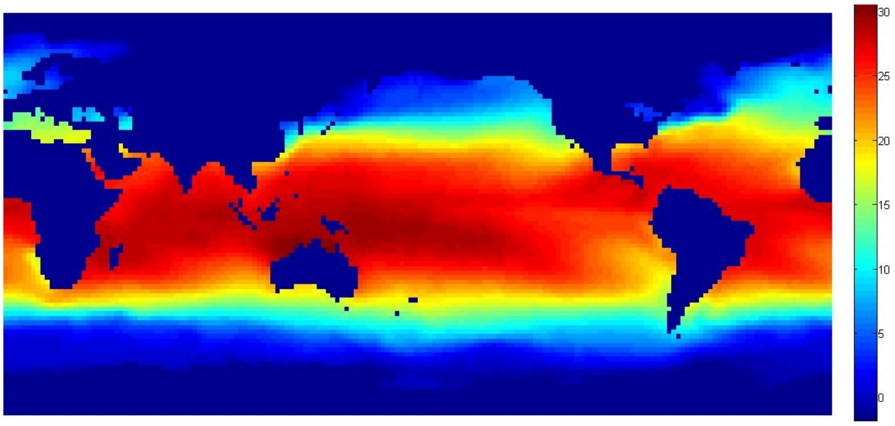

Figure 2.3: ERSST SST data set at the beginning of 1970s

3 METHODS AND TOOLS

The methods used in the study are presented in order of complexity. The first method described is the correlation analysis, followed by the Empirical Orthogonal Function (EOF) and Canonical Correlation Analysis (CCA). At last the Input Variable Selection (IVS) methods are shown. The correlation analysis and EOF/CCA methods, are commonly used by climate researches in order to investigate the teleconnection between SST/SSTA and climate or hydro-meteorological processes. IVS is only recently studied in the literature. The goals of these methods are basically:

- compression of informations (EOF)
- assessment of the relationship between SST/SSTA and rainfall/streamflow (Correlation analysis, CCA, IVS)
- prediction of streamflow (CCA, IVS)

A framework of four steps is developed as follow:

1. *Correlation Analysis* to verify and assess the effects of SST on the hydro-meteorological variables of the basin;
2. *EOF* on SSTA to create new sets of variables of SSTA (indicators) that explain the largest amount of total variance of the SSTA process;
3. *CCA* between the new sets of EOFs variables of SSTA and selected EOFs of streamflow/rainfall anomalies to assess the relationship between the two input variables;
4. *Streamflow prediction models* using the new sets of EOFs variables of SSTA and streamflow anomalies at each station of the basin. Moreover IVS method gives an evaluations of the SSTA indicators.

3.1 CORRELATION ANALYSIS

The Pearson product-moment correlation coefficient r_i quantifies the linear association between the SST/SSTA of a grid square cell and the rainfall and streamflow series. Basically, the method constitutes a first approach to the analysis of climate data, in order to obtain results about the relationship of the variables.

The coefficient is defined as follows

$$r_i = \frac{\sum_{i=1}^{k}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{k}(x_i - \overline{x})^2 \sum_{i=1}^{k}(y_i - \overline{y})^2}} \tag{3.1}$$

where for k observations the vector $X_i = [x_1, x_2, \dots x_k]$ represents the SST/SSTA time series in a grid square cell, and vector $Y = [y_1, y_2, \dots, y_k]$ represents the weather series (rainfall or streamflow data) recorded in a station. The method assumes a linear relationship among X_i, the predictor, and Y, the predictand. The values are between -1 and +1 inclusive, where +1 is total positive correlation, 0 is no correlation, and 1 is total negative correlation. It is important to note that it is possible to obtain a statistically significant correlation simply by correlating two random number series and, as in many studies, the coefficient's significance should be assessed to be greater than 95% by means of Student's t test.

There are a number of possible advantages of using correlation analysis. Foremost, it is possible to cluster local or regional grid cells of SST/SSTA observing the correlation coefficient at each grid point. Local and regional clustered cells can be easily understood starting from their spatial information and also they can be easily related to physical processes.

3.2 EMPIRICAL ORTHOGONAL FUNCTIONS AND CANONICAL CORRELATION ANALYSIS

3.2.1 *Empirical Orthogonal Functions*

Empirical Orthogonal Function (EOF) analysis or Principal Component Analysis (PCA) is a technique used for describing large data sets efficiently. The technique was originally described by Pearson (1982) and Hotelling (1935). It was first used in meteorology by Lorenz (1956). In meteorology and oceanography EOF analysis is used to describe a large number of time series efficiently, such as gridded SST/SSTA data.
The basic idea is as follows. Consider m time series $X_i(t)(i = 1, 2, \dots, m)$, each time series having zero mean. In order to describe the variability of these time series as simply as possible, linear combinations of the m time series are selected to explain as much as possible of the variability of the m time series [*Jolliffe*, 2002].

Weighting factors p_i are selected so that the variance of the linear combination $\sum_{i=1}^{m}(p_i X_i(t))$ is maximized; i.e. p_i is selected by maximizing $\sum_t \sum_{i=1}^{m}(p_i X_i(t))$ where $\sum_t$ denotes summation over all times t

of the time series.
Notice that the maximum of the above quantity is always found by letting $p_i \to \infty$, so to make the problem well-defined p_i is normalized by requiring

$$\sum_{i=1}^{m} (p_i^2) = 1 \tag{3.2}$$

In practice, the time series $X_i(t), i = 1, 2, \dots, m$ often correspond to time series of a physical variable at a spatial locations i. For example, the i-th cell of a grid square of SST/SSTA or the i-th rainfall/streamflow station.

The constrained maximization described above is solved by method of Lagrange multiplier and is equivalent to the maximization of

$$\varepsilon = \sum_{t} \Big(\sum_{i=1}^{m} (p_i X_i) \Big)^2 + \lambda \Big(1 - \sum_{i=1}^{m} (p_i)^2 \Big) \tag{3.3}$$

where λ is a Lagrange multiplier.
The maximum must satisfy $\delta\varepsilon / \delta\lambda = 0$ and this implies the satisfaction of the condition (2.2). The maximum must also satisfy $\delta\varepsilon / \delta p_k = 0$ and this gives

$$S_{xx} \mathbf{p} = \lambda \mathbf{p} \tag{3.4}$$

where S_{xx} is a matrix having the value $\sum_t (X_k(t) X_i(t))$ in the k-th row and i-th column. It then follows the normalization condition (2.2) that

$$\sum_{t} \Big(\sum_{i=1}^{m} (p_i X_i) \Big)^2 = \lambda \tag{3.5}$$

Since S_{xx} is a real symmetric matrix, all its eigenvalues are real and there exist real eigenvectors $\mathbf{p}$ such that (2.2) is satisfied. As p_i and X_i are real, by (2.5) λ is non-negative and hence all eigenvalues of (2.4) are non-negative.

The eigenvector $\mathbf{p}^{(i)}$ is known as the i-th empirical orthogonal function (EOF) and describes a spatial pattern of weighting. The linear combination $\boldsymbol{\alpha}_i(t) = \mathbf{p}^{(i)} \cdot \mathbf{X}$ is known as the i-th principal component. The m mutually perpendicular EOF vectors $\mathbf{p}^{(i)}$ span an m-dimensional space, so for coefficients $\boldsymbol{\alpha}_i$ it could be written

$$\mathbf{X} = \sum_{i=1}^{m} (\boldsymbol{\alpha}_i(t) \mathbf{p}^{(i)}) \tag{3.6}$$

Furthermore, the linear combination of the time series that maximizes the variability is associated with the eigenvector $\mathbf{p}^{(1)}$ corresponding to the maximum eigenvalue ($\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m$). This eigenvector $\mathbf{p}^{(1)}$ is known as the first empirical orthogonal function. Other eigenvectors are ordered in a descending way, according to the value of the corrisponding eigenvalue.

EOFs are particularly useful when the time series $X_i(t)$ are influenced by a single large-scale process. For example, suppose there are records of very low-frequency sea level along an ocean boundary and that there is no significant wind along the coast. Since there is no significant flow into the boundary, the sea level is expected spatially constant along the boundary and have the same time variability. This can be descripted by one dominant EOF $\mathbf{p}^{(1)}$ with all elements equal [*Jolliffe*, 2002].
One useful property of the EOFs is that they are orthogonal in space, consequently the principal components $\boldsymbol{\alpha}$ are uncorrelated in time, i.e., if $i \neq j$ then

$$\sum_t (\boldsymbol{\alpha}_i(t)\boldsymbol{\alpha}_j(t)) = 0 \tag{3.7}$$

These properties allow EOF analysis not to show two or more processes simultaneously and meanwhile show the major processes separately through the first eigenvectors. Thanks to that, one of the main purpose of EOF analysis is data compaction and filtering.
It is usually performed to derive SST/SSTA data set in order to retain all the large scale spatial dependencies that may exist in the record, and eliminate the "noise" of smaller scale features that are not useful in the context of global scale general circulation patterns [*Sharma*, 2000].

Specifically, EOF is used to compress geophysical predictor and predictand data sets in both space and time. This reduction is accomplished by projecting the temporal variance onto uncorrelated orthogonal spatial patterns (eigenvectors) and associated time series (principal components). The eigenvector patterns accounting for large variances are, in general, considered physically meaningful and connected with important centers of action. On the other hand, the remaining modes accounting for smaller variances are regarded as statistically and physically insignificant (noise) [*Roswintiarti, Devdutta, and Raman*, 1998].

There are two different methods to determine significant principal component and thus determine the number of them to retain: Kaiser method and Horn's Parallel Analysis (PA).

- Kaiser method (1960) is the most known and utilized in practice. According to this rule, only the principal components that have eigenvalues greater than one are retained for interpretation.
- Horn (1965) proposes PA, a method based on the generation of random variables. PA compares the observed eigenvalues of EOF's principal components with those obtained from uncorrelated normal variables. From a computational point of view, PA implies a Monte Carlo simulation process, since 'expected' eigenvalues are obtained by simulating normal random matrix of data of identical dimensionality to the observed matrix of data (same number of variables and samples). Principal components with eigenvalues higher than simulated ones are retained, otherwise principal components with eigenvalues lower than those obtained by Monte Carlo simulation are considered no significant.

3.2.2 *Canonical Correlation Analysis*

The Canonical Correlation Analysis (CCA) finds a linear combination of m predictor that maximizes correlation with a linear combination of n predictand time series. Actually, the time series need not to be predictors and predictands; the theory applies to maximizing correlation between a linear combination of any appropriate m time series with a linear combination of any appropriate n time series. Mathematically, the m time series of predictors is written in the form of a time-dependent column vector $\mathbf{X}(\mathbf{t})$ of length m and the n time series be written as a time-dependent column vector $\mathbf{Y}(\mathbf{t})$ of length n.

CCA produces the CCA modes, each containing:

- Eigenvectors (loading pattern): **p** and **q** that are interpreted as indicators of the underlying physical processes;
- Eigenvalues that show the importance of the relationship between predictor and predictand;
- Amplitude times series for each predictand and predictor variables.

Then if **p** is a time-independent vector of length m and **q** a time-independent vector of length n,

$$\alpha(t) = \mathbf{p} \cdot \mathbf{X} \tag{3.8}$$

and

$$\beta(t) = \mathbf{q} \cdot \mathbf{Y} \tag{3.9}$$

are linear combinations of the m and n time series, respectively. CCA finds **p** and **q** so that the correlation of $\alpha(t)$ and $\beta(t)$ is maximized.

$$r = \frac{\sum_t (\boldsymbol{\alpha}(\mathbf{t})\boldsymbol{\beta}(\mathbf{t}))}{\sqrt{\sum_t (\boldsymbol{\alpha}(\mathbf{t}))^2 \sum_t (\boldsymbol{\beta}(\mathbf{t}))^2}} \tag{3.10}$$

where the summation for t is over the length of the time series. Notice that if **p** or **q** are vectors describing optimal linear combinations and if σ is some scalar, then $\sigma\mathbf{p}$ or $\sigma\mathbf{q}$ are also optimal vectors since r in (2.10) remains unchanged.
In order to make **p** and **q**, a normalization is required

$$\sum_t (\boldsymbol{\alpha}^2(\mathbf{t})) = 1 \qquad and \qquad \sum_t (\boldsymbol{\beta}^2(\mathbf{t})) = 1 \tag{3.11}$$

And the optimal correlation is

$$r = \sum_t (\boldsymbol{\alpha}(\mathbf{t})\boldsymbol{\beta}(\mathbf{t})) \tag{3.12}$$

from which one can deduce that

$$r = \mathbf{p}^{\mathbf{T}} S_{XY} \mathbf{q} \tag{3.13}$$

So that the problem of maximizing r can be written in the Lagrange multiplier form and solved by derivative respect to Lagrange multipliers and setting the result equal to zero.
Let $\mathbf{p}^{(j)}$ and $\mathbf{q}^{(j)}$ be the j-th eigenvector pair and let the corresponding linear combinations be

$$\boldsymbol{\alpha}_j(t) = \mathbf{p}^{(j)} \cdot \mathbf{X} \tag{3.14}$$

and

$$\boldsymbol{\beta}_j(t) = \mathbf{q}^{(j)} \cdot \mathbf{Y} \tag{3.15}$$

A key property of the $\boldsymbol{\alpha}_j(t)$ or $\boldsymbol{\beta}_j(t)$ is that for $i \neq j$, $\boldsymbol{\alpha}_i(t)$ is not correlated with $\boldsymbol{\alpha}_j(t)$ and that $\boldsymbol{\beta}_i(t)$ is not correlated with $\boldsymbol{\beta}_j(t)$. Moreover for $i \neq j$ the correlation of $\boldsymbol{\alpha}_i(t) = \mathbf{p}^{(i)} \cdot \mathbf{X}$ with $\boldsymbol{\beta}_j(t) = \mathbf{q}^{(j)} \cdot \mathbf{Y}$ is zero.

CCA can also be viewed as a special form of EOF analysis, where the correlation structure between predictor and predictand datasets is described more completely with each successive CCA mode. It is used combining the CCA with EOF analysis. Prior to conducting the CCA, the standardized predictor and predictand data are separately condensed using EOF analysis, in order to reduce the number of original variables to fewer essential variables [*Barnston and Smith*, 1996].

3.2.3 *Prediction using CCA*

CCA is usefull to make predictions. Once CCA modes are obtained, predictions are given by using the amplitude time series of predictor (or predictand) to make predictions of predictor (or predictand). Suppose to estimate $\mathbf{Y}$ a time Δt into the future given $\mathbf{X}$ up to time t_*. It is possible to do that by estimating each component $Y_i(t_* + \Delta t)$ of $\mathbf{Y}(t_* + \Delta t)$ as the sum of M terms

$$\widehat{Y}(t_* + \Delta t) = \sum_{j=1}^{M} (\gamma_{ij} \boldsymbol{\alpha}_j (t_*)) \tag{3.16}$$

where the $\boldsymbol{\alpha}_j(t_*)$ is associated with the canonical correlation between $\mathbf{X}(t)$ and $\mathbf{Y}(t+\Delta t)$ instead of $\mathbf{X}(t)$ and $\mathbf{Y}(t)$. The coefficient γ_{ij} is chosen so that the fit is as good as possible in the least-squares sense by minimizing

$$\sum_t \left[Y_i(t+\Delta t) - \sum_{j=1}^{M} (\gamma_{ij} \boldsymbol{\alpha}_j (t)) \right]^2 \tag{3.17}$$

where the sum is taken over all times t for which $Y_i(t+\Delta t)$ and $\alpha_j(t)$ are available. Differentiating the above expression with respect to γ_{ij} and using the orthogonality of the α_j gives

$$\gamma_{ij} = \sum_t (Y_i(t+\Delta t) \boldsymbol{\alpha}_k (t) \tag{3.18}$$

In summary, if we have $\mathbf{X}$ up to time t_*, then by CCA between $\mathbf{X}(t)$ and $\mathbf{Y}(t+\Delta t)$, λ_{ij} can be estimated using known data up to time t_*. Since the $\mathbf{p}^{(j)}$ are known from CCA, at the time t_*, it can be calculated

$$\boldsymbol{\alpha}_j(t_*) = \mathbf{p}^{(j)} \cdot \mathbf{X}(\mathbf{t}_*) \tag{3.19}$$

and hence estimate the prediction $\widehat{\mathbf{Y}}(t_* + \Delta t)$.

3.3 INPUT VARIABLE SELECTION

Input Variable Selection (IVS) method is used to assess the relationship between SST/SSTA and rainfall/streamflow. Differently from the previous methods, the IVS method used, does not assume a linear relationship among the predictor and the predictand. For this reason it is an useful method to study relationships between variables such as SST/SSTA and rainfall/streamflow, that are ruled by complex phenomena.

3.3.1 *Overview*

The IVS problem is defined as the task of appropriately selecting a subset of s variables from an initial candidate set which comprises the set of all potential inputs to a model (i.e. candidates). The goal is to correctly identify the subset of input variables that collectively possess the largest amount of information about the system being modeled, without including irrelevant input variables [*Guyon and Elisseeff*, 2003; *May et al.*, 2008; *Hejazi and Cai*, 2009; *Galelli et al.*, 2014].

Defining an appropriate subset of input variables requires assessing the effect the choice of input variables ultimately has on the performance of the model that is either incorrectly over-specified or under-specified. An inaccurate model results when the input set is under-specified, as the selected variables do not fully describe the observed behavior within the system under consideration. On the other hand, the inclusion of input variables that are either irrelevant or redundant (i.e. over-specification) increases the size of the model. This not only adds processing time for model development and deployment, but it also adds noise, rather than information, to the model inputs and thus reduces accuracy [*Guyon and Elisseeff*, 2003]. Given these considerations, an appropriate set of model inputs is considered to be the smallest set of input variables required to adequately describe the observed behavior of the system [*May et al.*, 2008].

3.3.2 *Input Variable Selection techniques*

Input Variable Selection methods can be distinguished in wrappers (model-based approach) and filters (model-free approach) [*Guyon and Elisseeff*, 2003]. A brief description of the two approaches is provided below.

3.3.2.1 *Wrappers*

The model-based approach is based on the idea of calibrating and validating a number of models with different sets of inputs, and selecting the set that ensures the best model performance. Wrappers essentially treat the selection of inputs as an overall optimization of the model structure: the candidate inputs are evaluated in terms of prediction accuracy of a preselected underlying model.

Implementation of wrappers can be achieved in several ways, such as the combination of global optimization techniques (e.g. evolutionary optimization) with data-driven modeling (e.g. Artificial Neural Networks) in order to define the subset of input variables that maximizes the underlying model performance. The main drawback of this approach stands in its computational requirements, as a large number of calibration and validation runs must be performed to single out the best combination of inputs. This means that wrappers do not scale well when dealing with large data sets. Moreover, the input selection result depends on the predefined model class and architecture. Thus, on one hand, model-based approaches generally achieve better performances since they are tuned to the specific interactions between the model class and the data; but, on the other hand, the optimality of a selected set of inputs obtained with a particular model is not guaranteed for another one, and this restricts the applicability of the selected set [*Galelli and Castelletti*, 2013b].

3.3.2.2 *Filters*

In contrast to the model-based wrapper approach, in model-free filter techniques the variable selection is directly based on the information content of the candidate input data set as measured by statistical dependence between the candidates and the output variable [*Guyon and Elisseeff*, 2003]. Not having to deal with model class and calibration within the IVS problem, this approach not only leads to improved computational efficiency, but also results in input sets with wider applicability to different model architectures. However, the performance of IVS filters is largely dependent on the statistical dependency measure that is used [*Guyon and Elisseeff*, 2003]. Moreover, the significance measure is generally monotonic and, thus, without a predefined cutoff criterion, the commonly used algorithms tend to select very large subsets of input variables, with high risk of redundancy [*Galelli and Castelletti*, 2013b].

3.3.3 *Iterative Input variable Selection*

The tree-based Iterative Input variable Selection (IIS) is a novel hybrid approach, that incorporates some of the features of model-based approaches into a fast model-free method able to handle very large candidate input sets. The optimal subset is incrementally built using the information content of the data with a ranking-based procedure and then validated by a model-based forward selection process [*Galelli and Castelletti*, 2013b].

Description of the IIS algorithm

The IIS algorithm is embedded into a stepwise forward selection approach that consists of three main steps (see figure 3.1):

Step 1. Given s candidate inputs, the IIS algorithm runs an Input Ranking (IR) algorithm to sort the s candidate inputs according to a non-linear statistical measure of significance. Each candidate input is

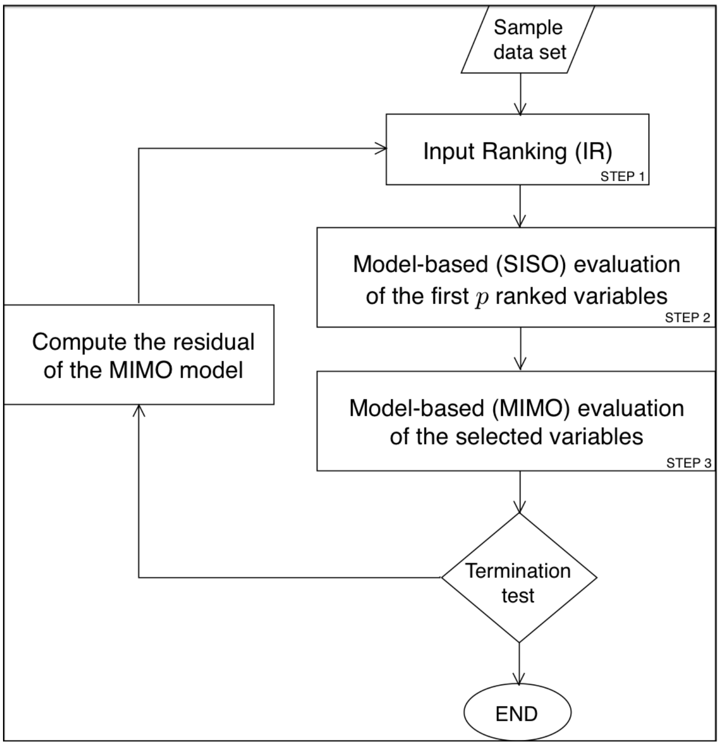
Sample data set
Input Ranking (IR)
STEP 1
Model-based (SISO) evaluation of the first p ranked variables
STEP 2
Model-based (MIMO) evaluation of the selected variables
STEP 3
Compute the residual of the MIMO model
Termination test
END

Figure 3.1: Flowchart of the IIS algorithm. Source: *Galelli and Castelletti* [2013b].

scored by estimating its contribution, in terms of variance reduction, to the building of the underlying model of the output y. In principle, the first variable in the ranking should be the most significant in explaining the output; in practice, having several potentially significant but redundant inputs, makes their contribution to the output explanation equally partitioned. This means that the most relevant input variables might not be listed in the very top positions. To reduce the risk for misselection, the first p variables in the ranking are individually evaluated in the following step.

Step 2. The relative significance of the first p-ranked variables is assessed against the observed output y. To this end, p Single Input-Single Output (SISO) models are identified with an appropriate Model Building (MB) algorithm and compared in terms of a suitable distance metric (e.g. mean-squared error) between the output y and each SISO model prediction[1]. The best performing input among the p consid-

[1] The evaluation of the chosen distance metric follows a k—fold cross-validation approach: the training data set is randomly split into k mutually exclusive subsets of equivalent size, and the MB algorithm is run k times. Each time the underlying model is validated on one of the k folds and calibrated using the remaining k–1 folds. The estimated prediction accuracy is then the average value of the metric over the k validations. The k—fold cross validation is aimed at estimating the ability of the model to capture the behavior of unseen or future observation data from the same underlying process, and, as such, it minimizes the risk of over-fitting the data.

ered is added to the set of the variables selected to explain y.

Step 3. A MB algorithm is then run to identify a Multi Input-Single Output (MISO) model mapping the variables so far selected into the output y.

The algorithm is not stopped until either the best variable returned by the IR algorithm is already in the set, or the performance of the underlying model does not significantly improve. At each iteration, the process is repeated using the residuals as the new output variable in steps 1 and 2: the reevaluation of the ranking on the model residuals every time a candidate variable is selected ensures that all the candidates that are highly correlated with the selected variable, and thus may become useless, are discarded. This strategy reinforces the SISO model-based evaluation in step 2 against the selection of redundant variables and is independent of the MB and IR algorithms adopted [*Galelli and Castelletti*, 2013b].

Extra-Trees

The IIS algorithm employs Extremely randomized Trees (ET) as underlying model family: they are a non-parametric tree-based regression method, originally proposed by Geurts, Ernst, and Wehenkel [2006]. Tree-based regressors are based on the idea of decision trees, which are tree-like structures composed of decision nodes, branches, and leaves, which form a cascade of rules leading to numerical values. Each tree is obtained by first partitioning at the top decision node, with a proper splitting criterion, the set of the input variables into two subsets, thus creating the former two branches. The splitting process is then repeated in a recursive way on each derived subset, until some termination criterion is met, e.g., the numerical values belonging to a subset vary just slightly or only few elements remain. When this process is over, the tree branches represent the hierarchical structure of the subset partitions, while the leaves are the smallest subsets associated to the terminal branches. Each leaf is finally labeled with a numerical value.

The Extra-Trees based algorithm is an ensemble method, which means that (using a top-down approach) it grows an ensemble of M trees. For each tree, the decision nodes are split using the following rule: K alternative cut directions (i.e. input variables x_i with $i = 1, \ldots, K$ candidate to be the argument of the node splitting criterion) are randomly selected and, for each one, a random cut-point is chosen; the variance reduction is computed for each cut direction, and the cut direction maximizing this score is adopted to split the node. The algorithm stops partitioning a node if its cardinality is smaller than a predefined threshold n_{min} and the node is a leaf in the tree structure. Each leaf is assigned with a value, obtained as the average of the outputs y associated to the inputs that fall in that leaf. The estimates produced by each single tree are finally aggregated with an arithmetic average over the ensemble of M trees, and the associated variance reduction of the ensemble is calculated.

The rationale behind the approach is that the combined use of randomization and ensemble averaging provides more effective variance reduction than other randomized methods, while minimizing the bias of the final estimate [*Geurts, Ernst, and Wehenkel*, 2006]. Although based on the construction of an ensemble of trees, the approach is still computationally efficient because the splitting rule adopted is very simple, if compared to other splitting rules that locally optimize the cut points, as, for example, the one in classification and regression trees [*Breiman et al.*, 1984]. A detailed analysis of the sensitivity of Extra-Trees performance to the tuning of M, K and n_{min}, with an application to a streamflow modeling problem, is presented in Galelli and Castelletti [2013a].
ET have been recognized to perform particularly well in characterizing strongly non-linear relationships and to provide more flexibility and scalability compared to parametric models such as Artificial Neural Networks [*Geurts, Ernst, and Wehenkel*, 2006]. In addition, the particular structure of Extra-Trees can be exploited to infer the relative importance of the input variables and to order them accordingly. This means that Extra-Trees can directly be used as an IR procedure.

The Input Ranking algorithm

Thanks to the particular structure of Extra-Trees, it is possible to rank the importance of the s input variables in explaining the output behavior. The ranking-based procedure has two main advantages:

1. It does not require any assumption on the statistical properties of the input data set (e.g., Gaussian distribution) and, thus, can be applied to any sort of sample;
2. It does not rely on computationally intensive methods (e.g., bootstrapping) to estimate the information content in the data and, thus, is generally faster and more efficient.

This approach is based on the idea of scoring each input variable by estimating the variance reduction it can be associated with by propagating the training data set over M different trees composing an ensemble.

Part II

PRACTICE

4 THE RED RIVER BASIN, VIETNAM

This chapter describes the study site: the Red River Basin in the north of Vietnam. The Red River Basin is the second largest of Vietnam after the Mekong River Basin. It is located between 20°-25.30°N of latitude, and 100°E-107.10°E of longitude, with a total area of approximately 169 000 km^2, of which 48% is in China, 51% in Vietnam, and the rest in Laos (as shown in figure 4.1) [*Quach*, 2011].

4.1 PHYSICAL SYSTEM

The Red River originates at the confluence of three main upstream tributaries, all originate from China: Lo, Thao and Da River. Even though the catchment areas of the Da and Thao River basins are almost the same, the Da River contributes 42%, while the Thao River contributes only 19% of total flow to the Red River. The contribution of the Lo River, the smallest one, is 25.4%. Properly the Da River, in the eastern part of the basin, pours first in the Thao River, and then, downstream Viettri, the Thao River and Lo River flows into Red River.

Lo River: it is 470 km long and its catchment is 39 000 km^2. The basin goes from being mountainous at the border with China, to becoming flat further downstream.

Thao River: it is 843 km long and its catchment is 51 800 km^2. It rises in a mountainous region of the Yunnan province, China, and its course is remarkably straight.

Da River: it is 900 km long and its catchment is 52 900 km^2. It also rises in China, in the south-western part of the Yunnan province, at an altitude of 2400 m a.s.l.

4.1.1 *Climate*

The basin is located in the South Asian Monsoon (SAM) region, over 10°-30°N and 70°-110°E, whose synoptic system is mainly of tropical nature. The whole basin is characterized by two distinguished seasons: wet season from May to October and dry season from November to April of the following year. During the wet season, winds blowing from South-South East bring humid air masses to the basin resulting in high temperatures and heavy rainfall; on the other hand,

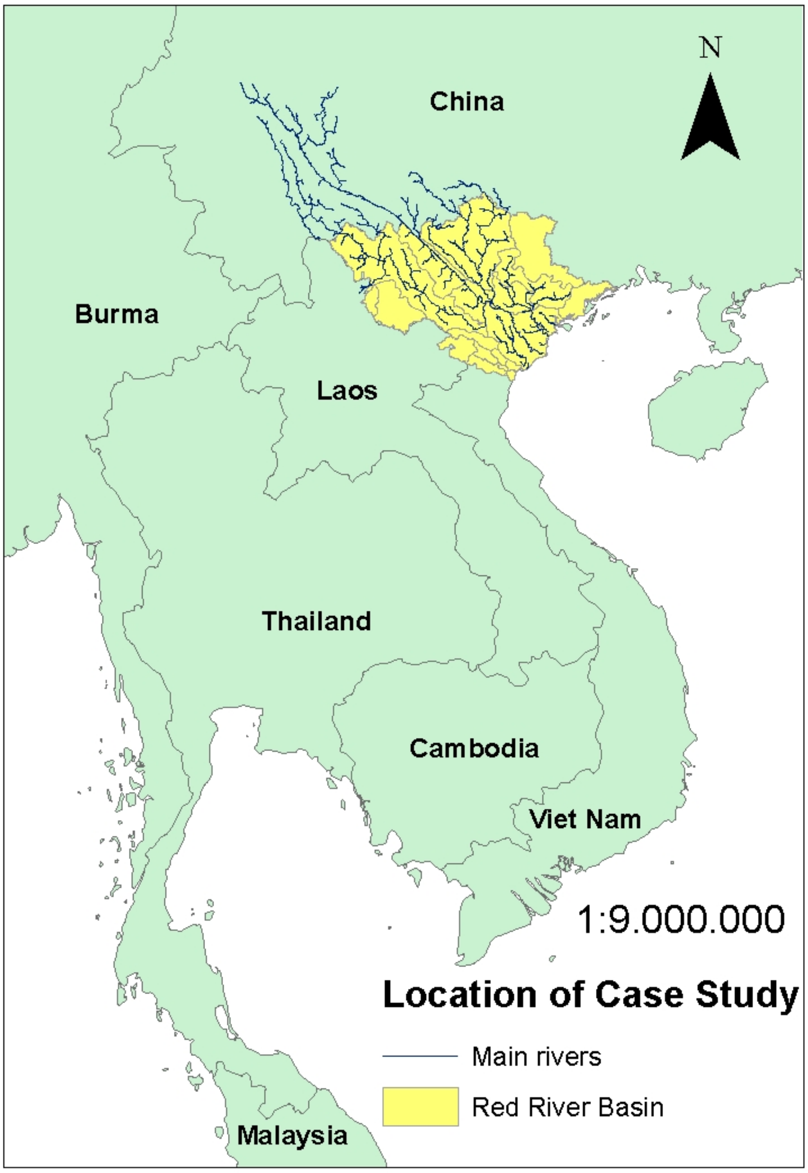

Figure 4.1: The Red River basin, Vietnam.

during the dry season, air circulation reverses direction to North East bringing dry air masses to the basin, inducing cooler weather and little rain.

4.1.2 *Rainfall*

Annual rainfall varies from 1200mm/year to 4800mm/year. As it can be seen in figure 4.2, the highest values of rainfall occurs during the wet season, from May to September, while the lowest rainfall is in from December to February in the dry season. About 80% of rainfall occurs during the wet season, with highest values occurring in July [*Quach*, 2011].

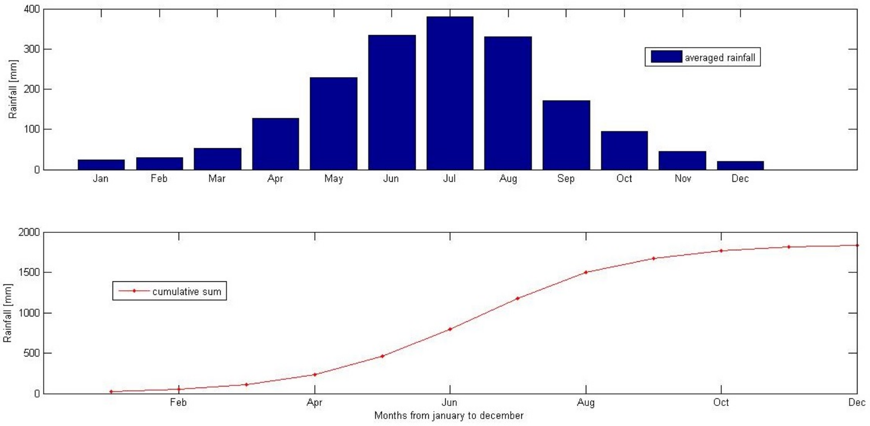

Figure 4.2: Rainfall trends in the Red River basin

4.1.3 *Streamflow*

Wet and dry seasons can be very well recognized also in the flow regime: flood (high flow) season goes from May to October, with the highest streamflows usually occurring in June-July, while low flow season is longer and goes from November to March. Because of this uneven distribution of rainfall, flows throughout the basin are unevenly distributed in time, causing floods and water-logging in the rainy season and water shortages in the dry season [*Quach*, 2011]. Streamflow trends in three stations located in three different river of the basin are shown in figure 4.3: Laichau and Hoabinh in the upper Da river, Chiemhoa in the Gam River (a tributary of Lo river).

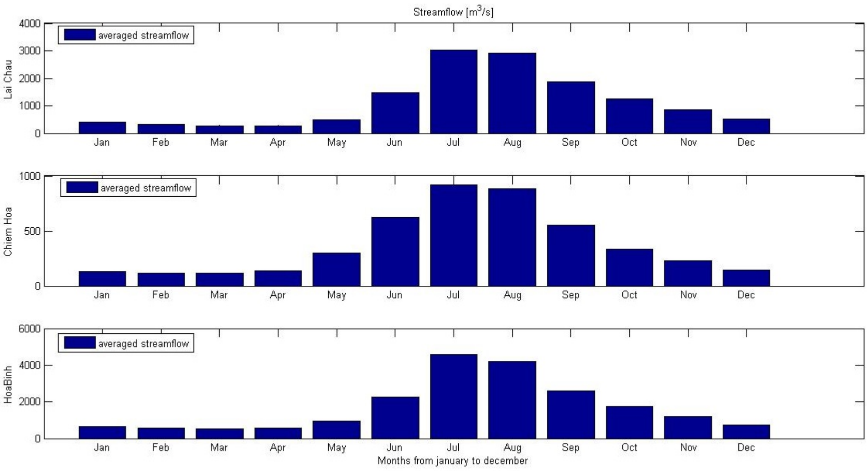

Figure 4.3: Streamflow trends in Laichau, Chiemhoa, Hoabinh stations

4.2 DATA DESCRIPTION

Rainfall and streamflow data come from a dense network of stations located in the basin and they made available by the Integrated and sustainable water Management of the Red River system (IMRR) research project. All the selected stations have been operated since the beginning of the 60s. Because of the different time of record in each station, the selection has been done to cover the longest time horizon with the largest possible number of stations. Monthly observations taken from January 1962 to April 2008, that consists in 47 years of records, are evaluated.

12 rainfall stations (see figure 4.3), located close to the Da River in the eastern part of the Red River basin, are utilized. They are clustered into 3 locations, going from the border of China to the downstream of the Da River: upper, middle and lower basin. The spatial aggregation is conducted by means. As shown in figure 4.4, rain distribution decreases southward, with the highest values recorded in the upper part of the basin. Table 4.1 summarizes the main characteristics of the stations considered in the river area. There are utilized 11 streamflow stations (see figure 4.3), spread in the whole Red River basin as summarized in Table 4.2.

Location	Stations	Latitude	Longitude	Period of record
Upper Basin	Muong Nhe	22,18°N	102,46°E	1960-2010
	Muong Te	22,36°N	102,83°E	1961-2011
	Lai Chau	22,05°N	103,15°E	1958-2011
	Sin Ho	22,35°N	103,25°E	1962-2011
Middle Basin	Tuan Giao	21,58°N	103,41°E	1961-2011
	Quynh Nhai	21,83°N	103,56°E	1961-2009
	Than Uyen	22,01°N	103,91°E	1962-2011
Lower Basin	Son La	21,33°N	103,90°E	1961-2011
	Yen Chau	21,05°N	104,28°E	1962-2011
	Phu Yen	21,26°N	104,65°E	1962-2011
	Moc Chau	20,85°N	104,63°E	1962-2011
	Hoa Binh	20,81°N	105,33°E	1957-2011

Table 4.1: Summary of the rainfall stations used for the Red River (see figure 4.5).

4.3 ENSO-RED RIVER TELECONNECTION

Beltrame and Carbonin [2013] studied the effect of ENSO on the Da River basin from May 1961 to April 2008. In their analysis the same rainfall stations of this study and the streamflow on station of HoaBinh are used. The classification of ENSO events is based on

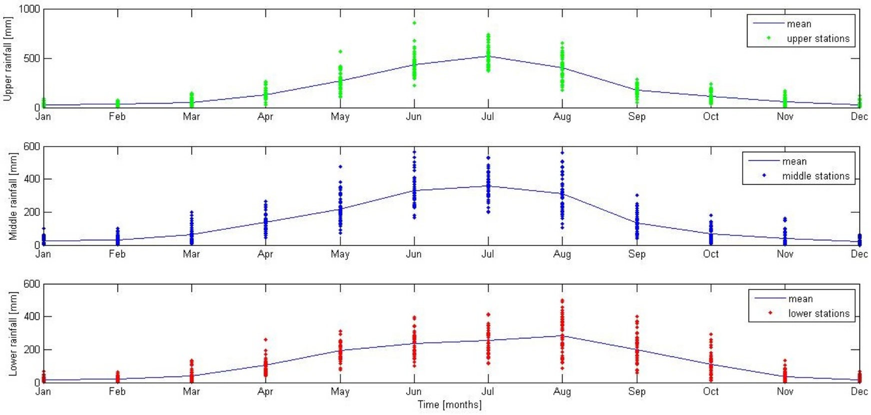

Figure 4.4: Metereological trends in the upper, middle, lower locations

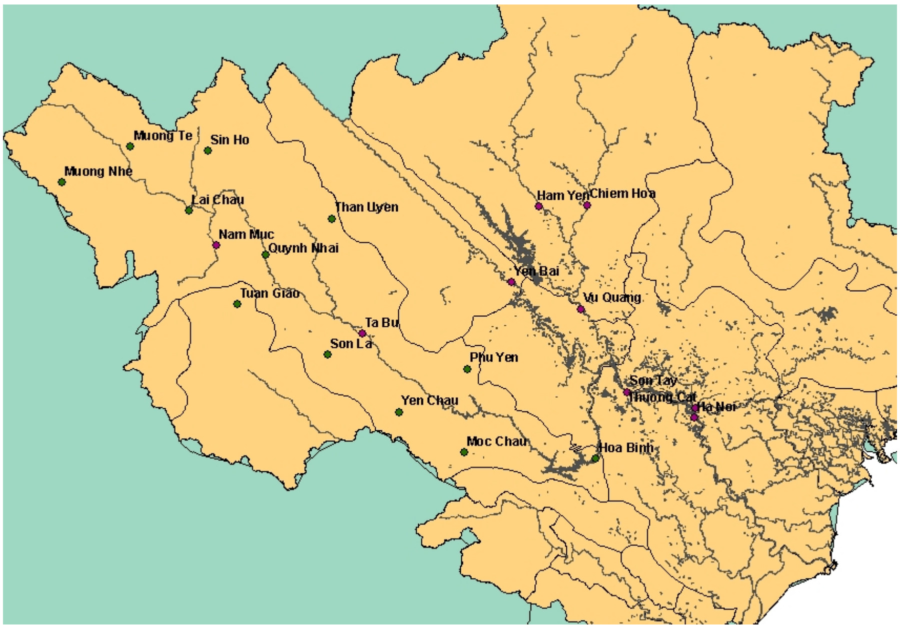

Figure 4.5: Map of the stations considered in the Red River basin

the Oceanic Niño Index (ONI), provided by NOAA: a-month-running mean of SST averaged in the Niño3.4 region.

Graphical analysis reveals that the ENSO influence on the Red River is quite weak for each data. The analysis on streamflow are consistent with findings of Räsänen and Kummu [2013] for the Mekong River, that is the potential impacts of ENSO are not much related to the maximum peak flow, it affects more the duration of the flood season, with a generally longer flood period during La Niña and conversely shorter during El Niño. Other results (see figure 4.6) show that sometimes the streamflow anomalies are oppositely aligned to ENSO phase, i.e. there are positive anomalies during La Niña and negative anomalies during El Niño, and this results could be interpreted as an influence of ENSO on streamflow anomalies. This is noticeable for the El Niño 1968-1969 event which causes a prolonged

River	Stations	Latitude	Longitude	Period of record
Da River	Lai Chau	22,06°N	103,16°E	1957-2011
	Nam Muc	21,87°N	103,29°E	1960-2011
	Ta Bu	21,43°N	104,05°E	1961-2011
	Hoa Binh	20,81°N	105,31°E	1956-2008
Thao River	Yen Bai	21,58°N	103,41°E	1956-2011
Lo River	Ham Yen	22,05°N	105,08°E	1960-2010
	Chiem Hoa	22,08°N	105,26°E	1959-2010
	Vu Quang	21,56°N	105,25°E	1957-2011
Red River	Son Tay	21,15°N	105,50°E	1956-2011
	Thuong Cat	21,06°N	105,86°E	1957-2011
	Ha Noi	21,01°N	105,85°E	1956-2011

Table 4.2: Summary of the streamflow stations used for the Red River (see figure 4.5).

period of negative monthly anomalies, reversed into positive anomalies by the subsequent La Niña 1970-1971. Other El Niño signals are visible for the events 1986-1987, when anomalies turn from positive to negative in a few months, 1992-1993 when persistent negative anomalies occur, and 2003-2004 similarly. Effects due to La Niña are found for the events 1995-1996, when increasingly positive anomalies are registered, and 1999-2000, when positive anomalies persist through many consecutive months. Results on rainfall are weaker than results on streamflow and difficult to seek when visualizing trajectories. The only considerable case reported in their study is related to rainfall in the lower basin, this could be explained considering that the influence of ENSO increases when moving southwards, as empirically demonstrated by Räsänen and Kummu [2013] for the nearby Mekong basin. Rainfall occurring in the summer peak can be variably high irrespective of El Niño or La Niña years [*Beltrame and Carbonin*, 2013]. However results reported do not show a strong relationship between ENSO and streamflow anomalies. Statistical analysis performed by box-plot, show appreciable difference in streamflow anomalies during El Niño and La Niña phases but not strong enough to be significant from a statistical point of view, and similarly to results on graphical analysis, rainfall statistical results aren't significant.

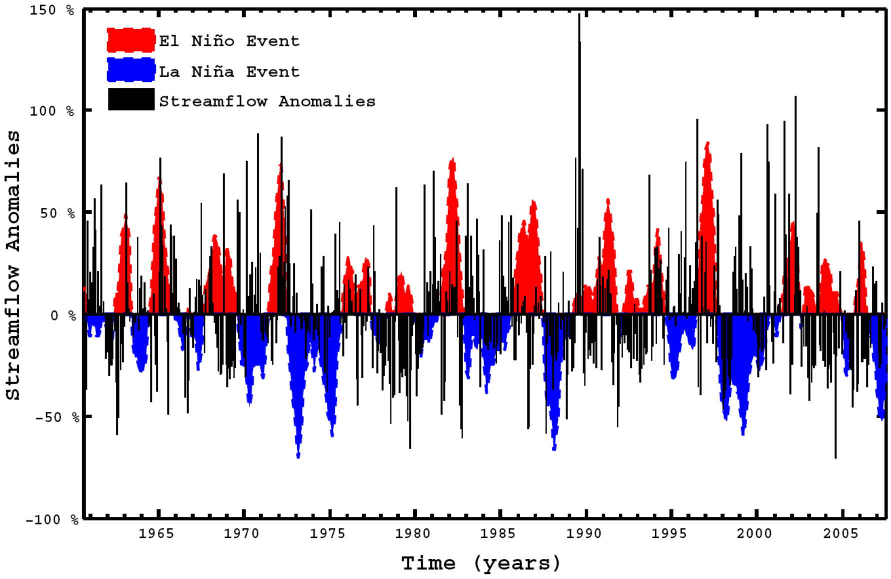

Figure 4.6: Sequence of ENSO events and Da River monthly streamflows in the period considered (1961-2007). ENSO events refer to the historical pattern of ONI. Streamflows represented in bars are monthly anomalies related to the respective monthly mean.

5 IMPACT OF SSTA ON HYDRO-METEOROLOGICAL PROCESSES

The scope of this chapter is to describe the analysis conducted to SST/SSTA in the Indian and Pacific Ocean and rainfall and streamflow in the Red River Basin. In particular the main task is to understand the existence of the teleconnection at first, and then to evaluate how and when oceans, through SST/SSTA, affect the hydro-meteorological processes in the Red River Basin.

Two principal ocean's areas are evaluated. A1 is centered to Indian Ocean and A2 is centered to Pacific Ocean. Indian Ocean and Pacific Ocean are selected in this work because they are the oceans closest to the basin, thus the teleconnection's effects are supposed to be very strong. The corect delimitation of the selected areas is summarized in Table 5 and displayed in figure 5.1 and figure 5.2.

Area	Latitude	Longitude	Ocean of reference
A1	40°N to -20°S	60°E to 150°E	Indian Ocean
A2	40°N to -20°S	110°E to 230°E	Pacific Ocean

Table 5.1: Areas of oceans evaluated

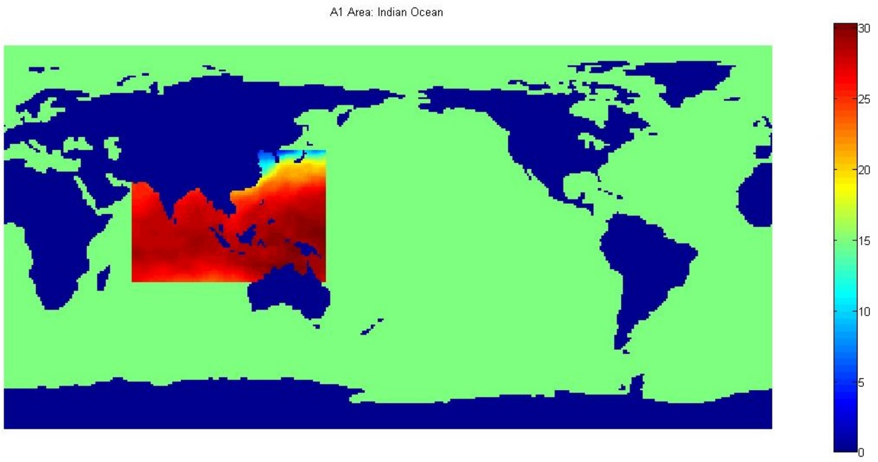

Figure 5.1: A1 Area: spatial location SST in January 1971

In next sections results obtained from Correlation Analysis, Empirical Orthogonal Function (EOF) and Canonical Correlation Analysis (CCA) are shown. Analysis are evaluated on SST, rainfall and stream-

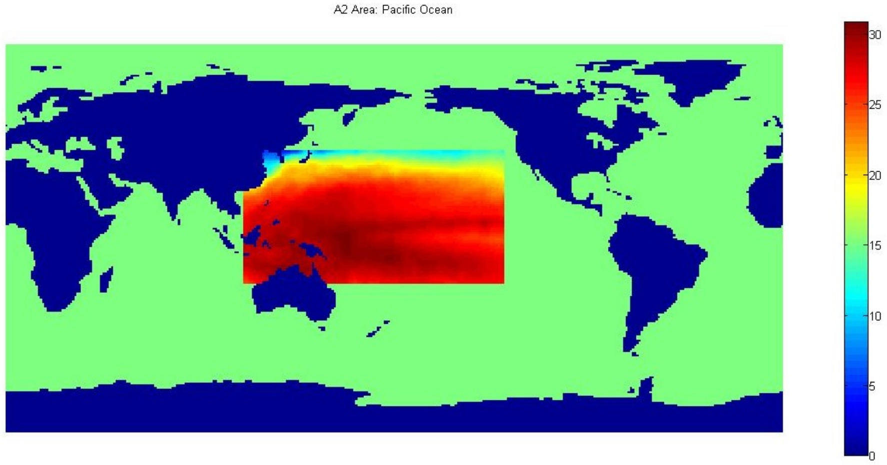

Figure 5.2: A2 Area: spatial location SST in January 1971

flow in stations in Red River Basin and also on their anomaly time series (SSTA, rainfall anomalies and streamflow anomalies). An anomaly time series is defined as the time series of deviations of a quantity from the long-term average. In practice the long-term monthly averages are removed from each monthly value of the time series. All data sets are monthly series from January 1962 to April 2008. The SST data set chosen is the ERSST (see section 1.3.2).

5.1 RESULTS AND DISCUSSION

5.1.1 *Correlation Analysis Results*

Correlation Analysis are done on each grid cell of SST over Indian Ocean/Pacific Ocean and rainfall/streamflow series.

Rainfall
First a correlation analysis is conducted between upper, middle and lower locations of rainfall in Da river and the SST at each grid point. The purpose of this first analysis is to show if a relationship among the two variables exists. Results confirm that SST is correlated with rainfall and the pattern of correlation is not dependent on the spatial location of the rainfall. In fact, similar results are obtained from correlation of SST and rainfall in upper, middle and lower basin. This suggests that the correlation does not change over the Da river (e.g. see figure 5.3). Correlation Analysis is done with monthly SST and monthly rainfall series with a lag time from 1 to 6 months. In figure 5.4 are displayed the correlation from a lag time of 1 month to a lag time of 3 months and in figure 5.5 from a lag time of 4 months to a lag time of 6 months. The positive correlation area move southward from lag time 4 to lag time 6 and this result could be caused by circulation

of mass of ocean water. It is noticeable that the ENSO area in the Pacific Ocean is always positive correlated with rainfall and the correlation is strong. In particular, the shape of ENSO area is stressed in the correlation map of 1 and 2 lag time. Also the Indian Ocean is noticeable in map correlation, in particular from a lag time of 2 to 3 months, there is a strong positive correlation as in ENSO area. These first results suggest that effectively there is a link among SST and rainfall, and ENSO area and Indian Ocean area play a key role. A p-value test is run to test the significance of the results. Confirming that most of the correlation maps are significant. In figure 5.6 significant p-values for a lag time from 1 to 3 months are shown, and similarly in figure 5.7 for lag time from 4 to 6 months. The grid cells with p-value<0.01 are reported in orange. The ENSO area is significant also in the p-value maps, this is clear in lag time 1, where the ENSO area is shaped from the rest part of the ocean.

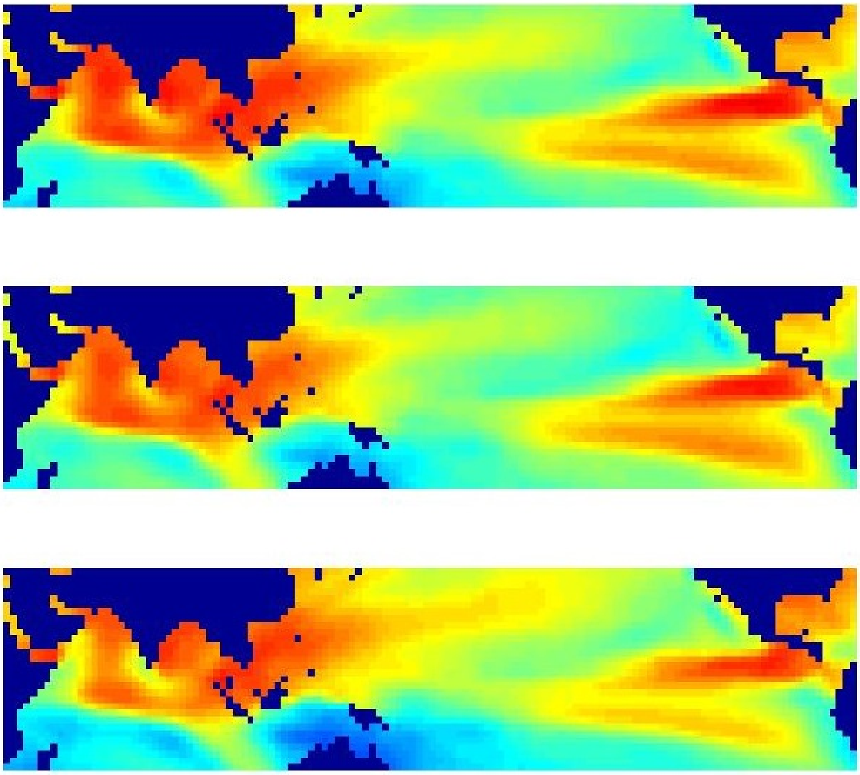

Figure 5.3: Correlation analysis among SST and rainfall in the three location at lag1. In order from top to bottom: upper, middle and lower basin. The correlation maps are similar.

Streamflow

The same correlation analysis is carried out for streamflow data (see section 3.1). The correlation are calculated for each grid point of SST with each streamflow station. The correlation maps are similar in ev-

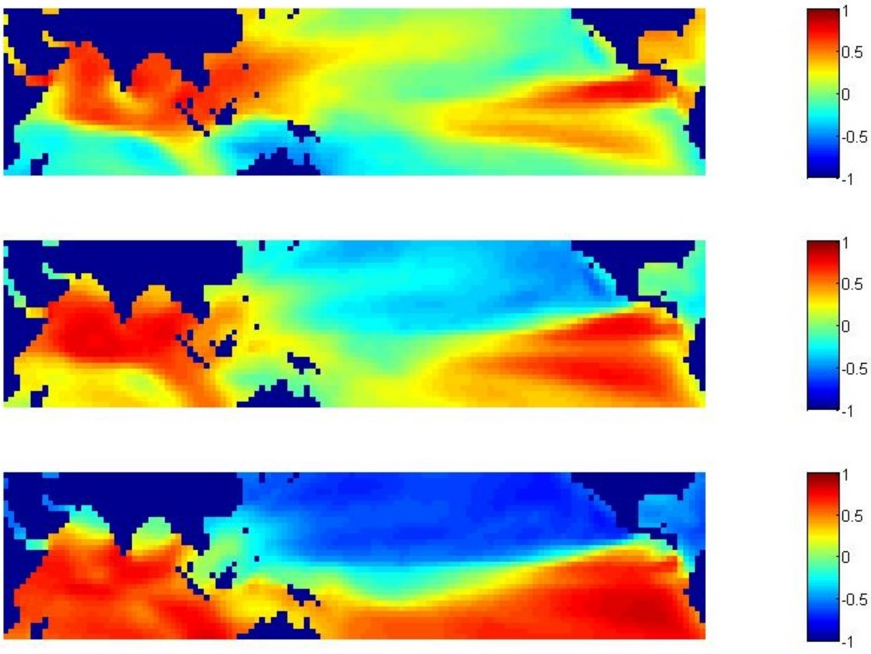

Figure 5.4: Correlation analysis among SST and rainfall in upper basin at different lag time: lag time 1 (above), lag time 2 (middle), lag time 3 (below). The lag time is in month.

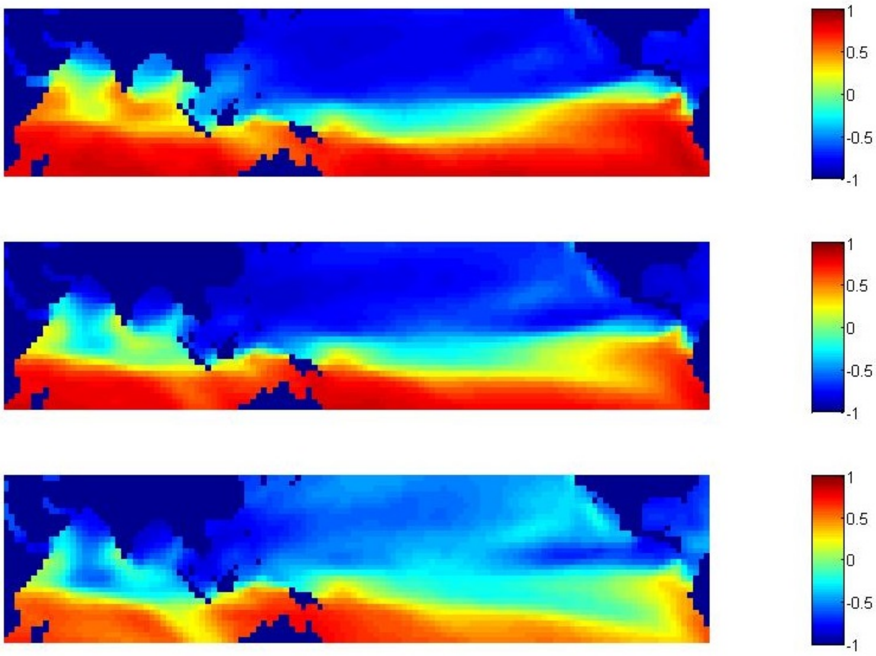

Figure 5.5: Correlation analysis among SST and rainfall in upper basin at different lag time: lag time 4 (above), lag time 5 (middle), lag time 6 (below). The lag time is in month.

ery station, no significant difference is observed, that means that the relationship among SST and streamflow is homogeneous in all rivers.

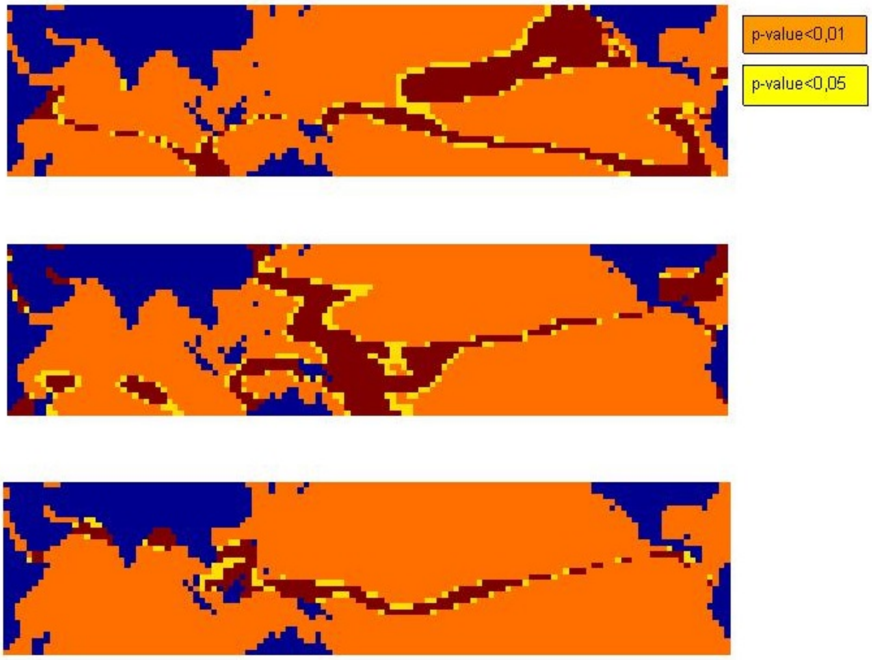

Figure 5.6: P-value test among SST and rainfall in upper basin at different lag time: lag time 1 (above), lag time 2 (middle), lag time 3 (below). Orange: p-value<0.01, yellow: p-value<0.05.

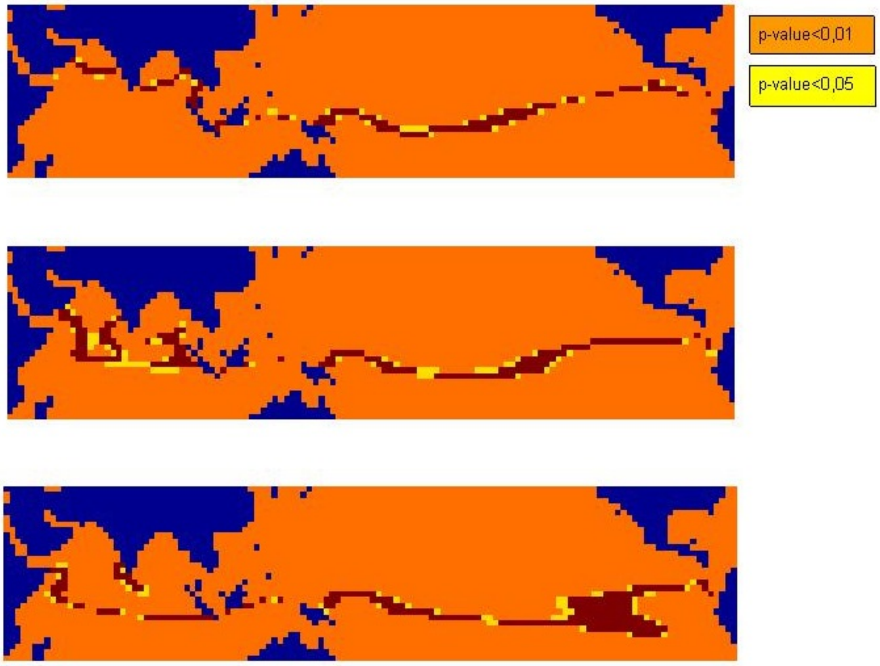

Figure 5.7: P-value test among SST and rainfall in upper basin at different lag time: lag time 4 (above), lag time 5 (middle), lag time 6 (below). Orange: p-value<0.01, yellow: p-value<0.05.

Moreover the patterns displayed in these maps are similar to those observed in the correlation maps of rainfall. In figure 5.8 the correlation maps of Chiem Hoa station are displayed from 1 lag time to 3 lag time. Similarly, in figure 5.9 the correlation maps are displayed from 4 to 6 lag time. All the most correlated area are significant, as shown in p-value maps figure 5.10 and figure 5.11. The ENSO area and Indian Ocean are strong and positive correlated with streamflow. The shape of ENSO area is well identified especially in 1 and 2 lag time. The strength of the ENSO area is less pronounced than in the rainfall's correlation maps, infact the relationship among hydrologic variables (streamflow) and SST is less directly correlated than the climatic (rainfall) one.

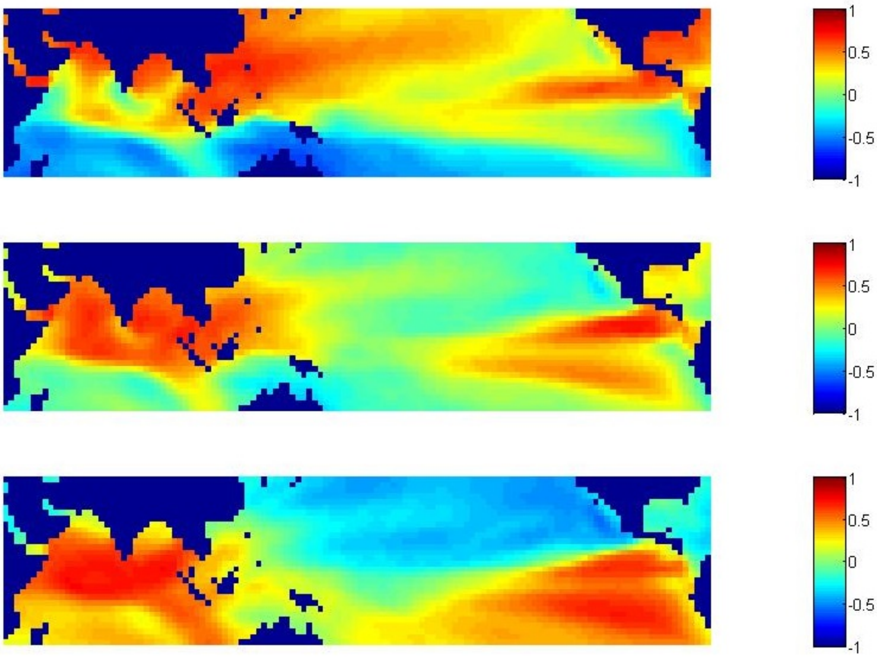

Figure 5.8: Correlation analysis among SST and streamflow in Chiem Hoa station at different lag time: lag time 1 (above), lag time 2 (middle), lag time 3 (below). The lag time is in month.

5.1.2 *Empirical Orthogonal Function Results*

Empirical Orthogonal Function are evaluated on every data set of variables: rainfall anomalies, streamflow anomalies, SSTA on A1 area and SSTA on A2 area. SSTA is used instead of using SST to assess the influence of the sea surface temperature no affected by seasonal variability.

Rainfall anomalies

EOF is computed on the 12 rainfall stations anomalies time series. The principal components are ordered in descending way of explained

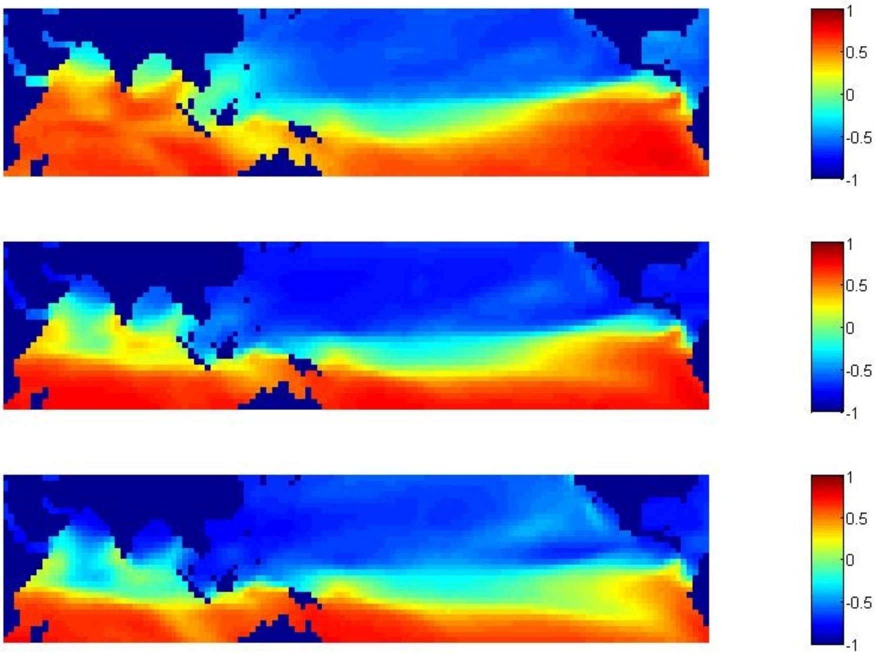

Figure 5.9: Correlation analysis among SST and streamflow in Chiem Hoa station at different lag time: lag time 4 (above), lag time 5 (middle), lag time 6 (below). The lag time is in month.

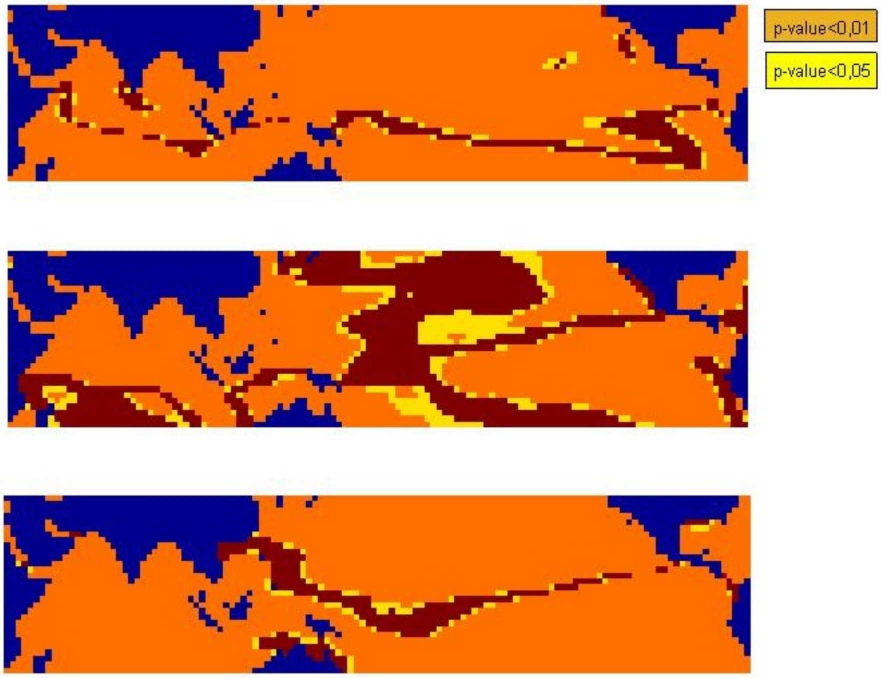

Figure 5.10: P-value test among SST and streamflow in Chiem Hoa station at different lag time: lag time 1 (above), lag time 2 (middle), lag time 3 (below). Orange: p-value<0.01, yellow: p-value<0.05.

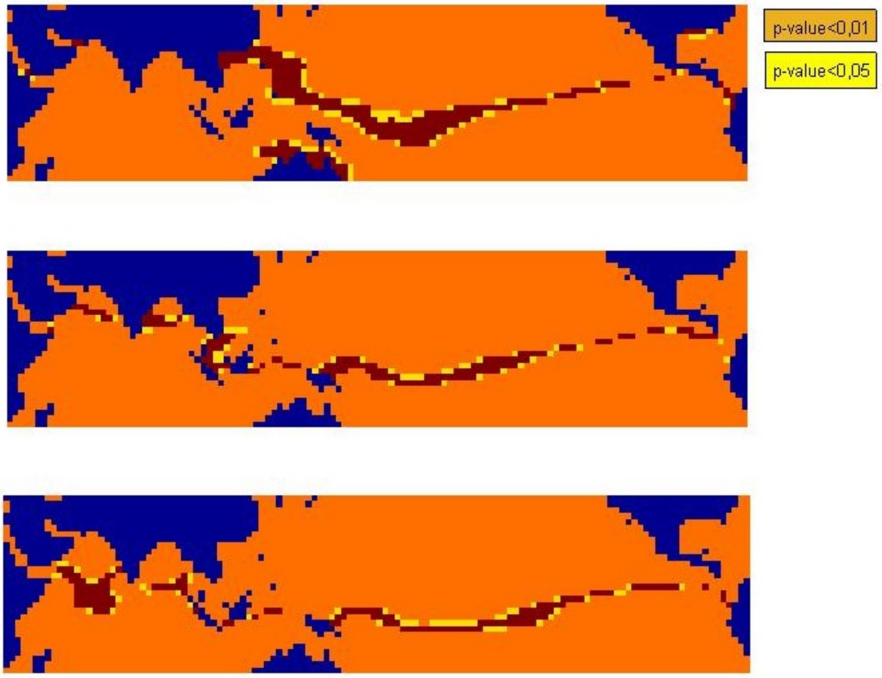

Figure 5.11: P-value test among SST and streamflow in Chiem Hoa station at different lag time: lag time 4 (above), lag time 5 (middle), lag time 6 (below). Orange: p-value<0.01, yellow: p-value<0.05.

variance, as shown in figure 5.12. The first two principal components are significant for the Horn method and the first three for the Kaiser method (see section 2.3.2). Following the Kaiser method, the first three principal components explain more than 60% of the total variance, such as, the first principal components explains the 40% of the process, the second less then 20% and the tirth around 10% of the total variance. Further principal components explain less of the process. Weighting factors pattern are different in each principal components and the most interesting patterns are displayed in the first two principal components. Stations are all weighted around 30% in the first principal components. The second principal components displayed a gradient (see figure 5.13). Weighting factors gradually increase going from north-west to south-east of the Da river. In the northest part stations are weighted around -30%, in the middle part weigths gradually increase from -10% to 30%. In the southest part, stations are positively weighted around 40%. The gradient in the weighting factors is explained by the "corridor-barrier" phenomenon [*Yungang, Daming, and Changqing*, 2008]. The Red River Basin is characterized by longitudinal montain ranges, major rivers and deep valleys, which had "corridor" function in south-north direction and "barrier" function in east-west direction on the trasportation of vapor. Thanks to the topography, masses of vapor are constrained to pass in the Da river valley from south-east and they lose their energy northward [*Yungang, Daming,* *and Changqing*, 2008].

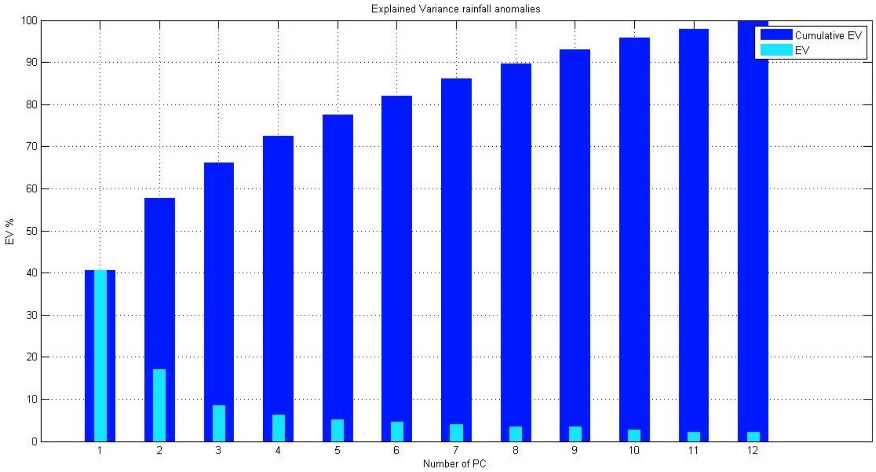

Figure 5.12: Explained Variance on the principal components of rainfall anomalies.

Streamflow anomalies
EOF on the 11 streamflow stations gives performance represented in figure 5.14. It is noticeable that only the first principal component explains more than 70% of explained variance, the second around 10% and the others less than 10%. As a consequence, only the first principal components is significant for the Horn method and the first two for the Kaiser method. As observed in the rainfall, also the first principal components of streamflow anomalies has the same weight (around 30%) for each station. The interpretation of the second principal component is more difficult than for the rainfall, because the streamflow stations are widespread all around the basin. In this case results are not displayed.

SSTA in area A1
EOF evaluates each SSTA grid cell as a variable that influence the behaviour of the whole SSTA area. Before running EOF methods, a spatial aggregation is computed on A1 and A2 areas. The grid cell are aggregated on a 4°x4°grid square in order to ensure a number of observations greater than that of the variables. Explained variance graph is shown in figure 5.15. The first 16 principal components are significant for Horn method and the first 20 for Kaiser method. Spatial patterns of weighting factors are displayed in maps and represent the role of each SSTA grid cell in the behaviour of the system. Since each principal component is orthogonal and so independent of another one, the weighting factors of each EOF display pattern independent. The first EOF (EOF1) spatial pattern, which has the highest value of explained variance (around 40%), shows the leading SSTA cells in the Indian Ocean. As displayed in figure 5.16, the cells most weighted

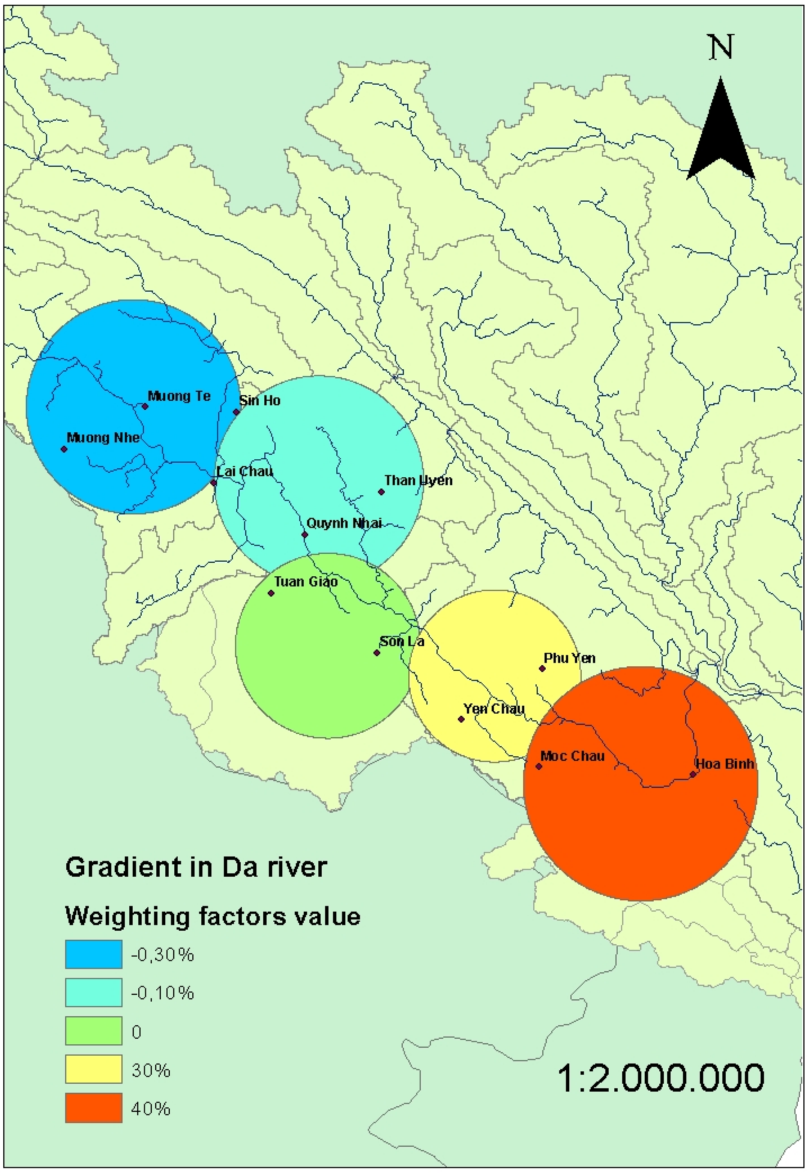

Figure 5.13: Spatial pattern of the second principal component of rainfall anomalies.

are clustered more in the Indian Ocean. In the EOF2 (more than 10% of explained variance and see figure 5.17) the ocean area above Australia close to Philippines is the more predominant, in EOF3 (around 5% of explained variance see figure 5.18) the leading ocean part is the sea close to Indonesia and in EOF4 (around 5% of explained variance and see figure 5.19) the weighted part is centered in China sea.

SSTA in area A2

EOFs of area A2 have values of explained variance lower than those

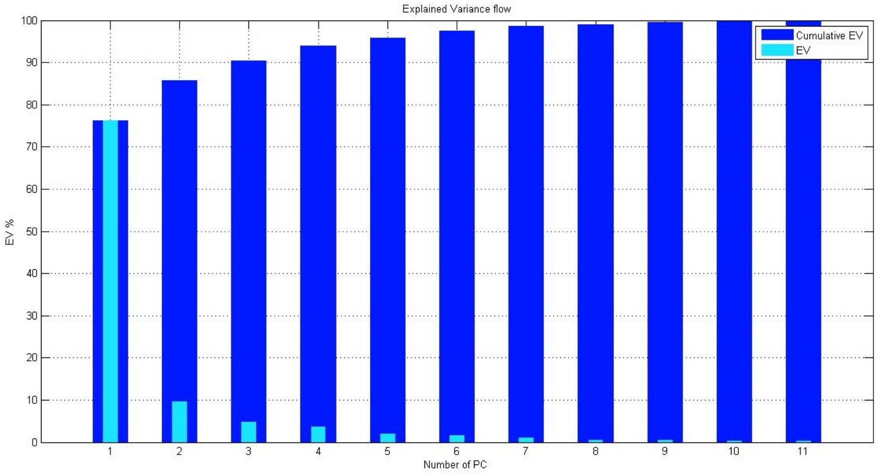

Figure 5.14: Explained Variance on the principal components of streamflow anomalies.

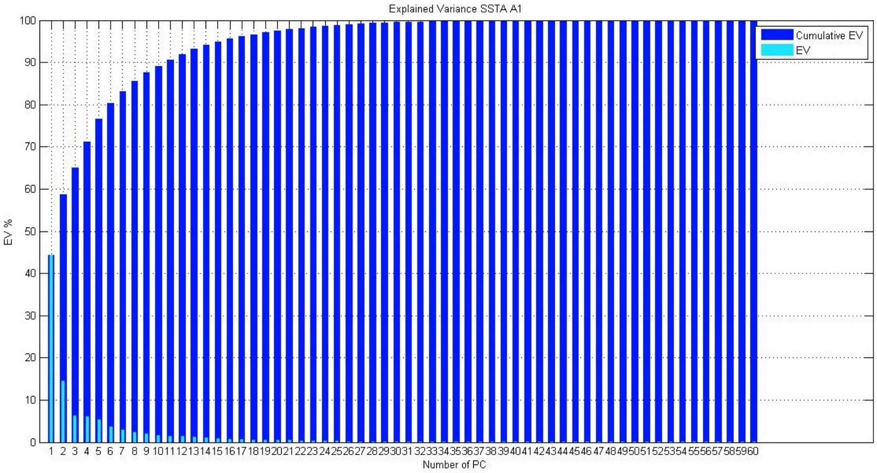

Figure 5.15: Explained Variance on the principal components of SSTA in area A1.

in area A1 (see figure 5.20). Infact the number of SST cells in area A2 is greater than in area A1. Kaiser method selects 29 principal components and Horn method 20 principal components. Spatial pattern of the first EOF (EOF1), which explains less than 30% of explained variance, underline the role of ENSO in area A2. In figure 5.21, only the cells in the ENSO area are positive weighted and that means that ENSO phenomenon is the leading behaviour in the area A2. Further EOFs can not be easily interpreted (see figure 5.22, figure 5.23 and figure 5.24). Weighted cells in EOF2 and EOF3 follow orthogonal directions and that is an effect of the independency of the EOFs. Cells in ENSO area are still weighted in EOF2 and EOF4. Because EOFs are independent, these results are not related to ENSO but they show the importance that ENSO area have for less important processes in the

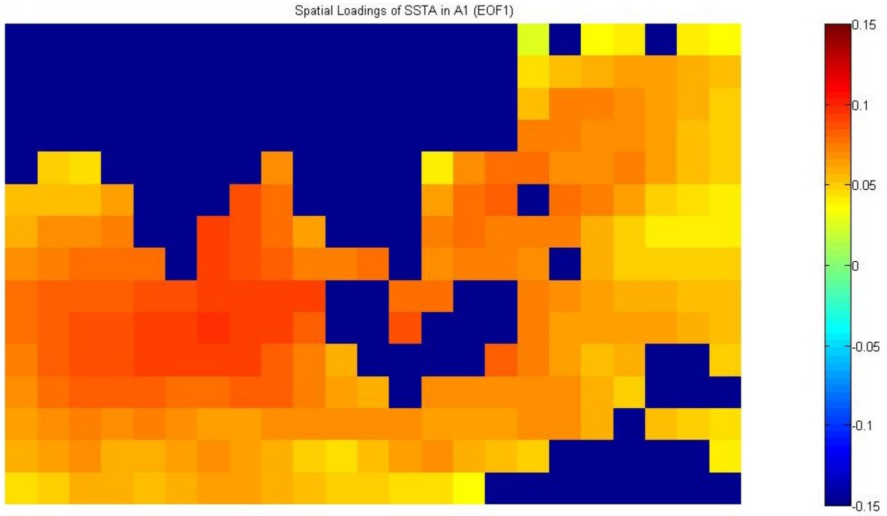

Figure 5.16: Spatial pattern of EOF1 in area A1. The weighting factors are higher in the Indian Ocean. Lands are displayed in blue, in particular Southeast Asia is displayed in the upper left part, North of Australia is displayed in bottom right (see figure 5.1 to have more spatial reference).

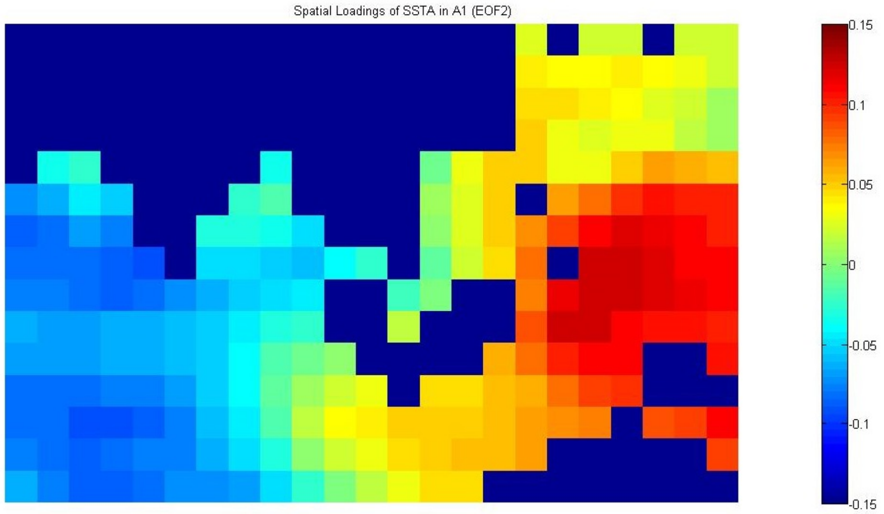

Figure 5.17: Spatial pattern of EOF2 in area A1. The weighting factors are higher in Philippines sea. Lands are displayed in blue, in particular Southeast Asia is displayed in the upper left part, North of Australia is displayed in bottom right (see figure 5.1 to have more spatial reference).

system.

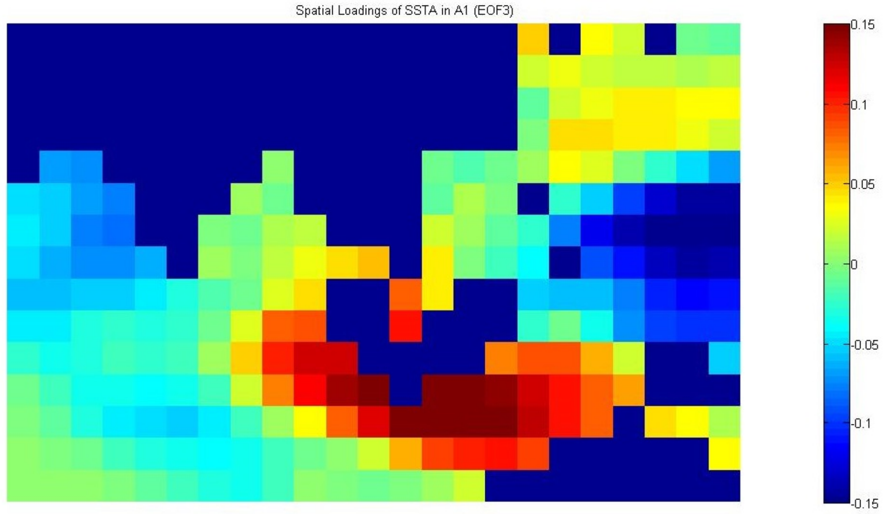

Figure 5.18: Spatial pattern of EOF3 in area A1. The weighting factors are higher in the Indonesia sea. Lands are displayed in blue, in particular Southeast Asia is displayed in the upper left part, North of Australia is displayed in bottom right (see figure 5.1 to have more spatial reference).

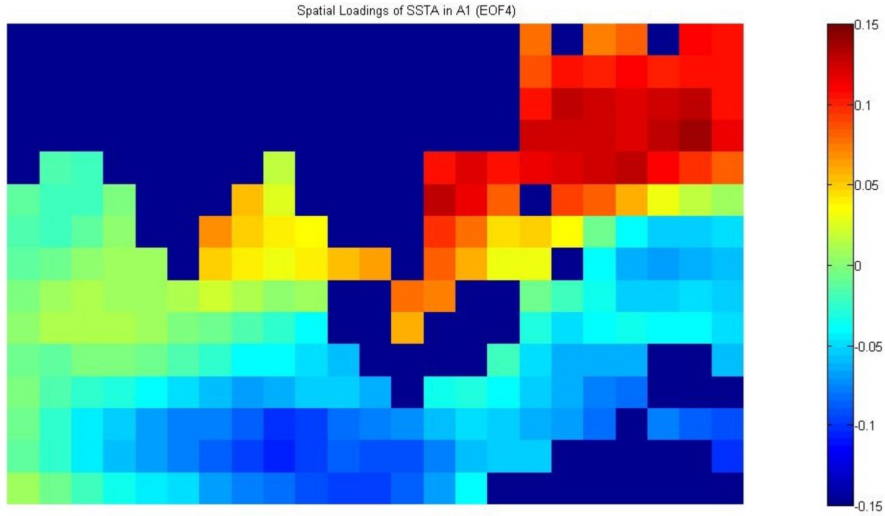

Figure 5.19: Spatial pattern of EOF4 in area A1. The weighting factors are higher in the China sea. Lands are displayed in blue, in particular Southeast Asia is displayed in the upper left part, North of Australia is displayed in bottom right (see figure 5.1 to have more spatial reference).

5.1.3 *Canonical Correlation Analysis Results*

CCA is computed with EOFs of SSTA as predictors and EOFs of rainfall/streamflow anomalies as predictands. The number of EOF retained to CCA is chosen using Kaiser method because it selects few more principal components than Horn method. The CCA experiments are

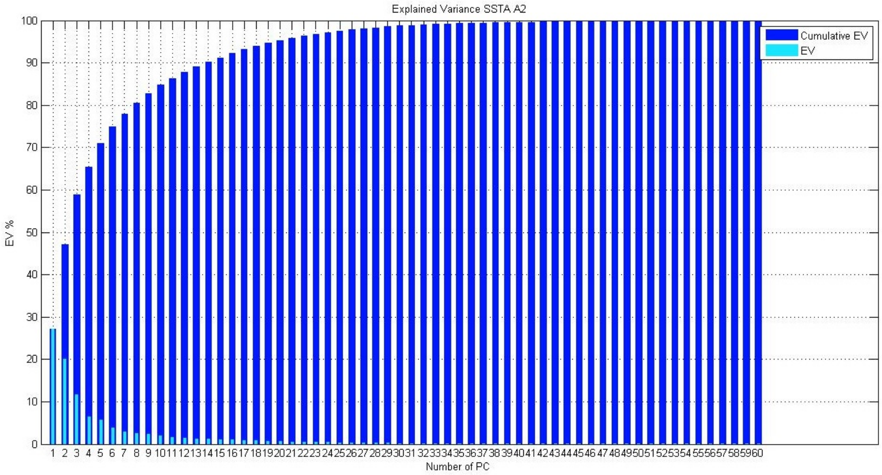

Figure 5.20: Explained Variance on the principal components of SSTA in area A2.

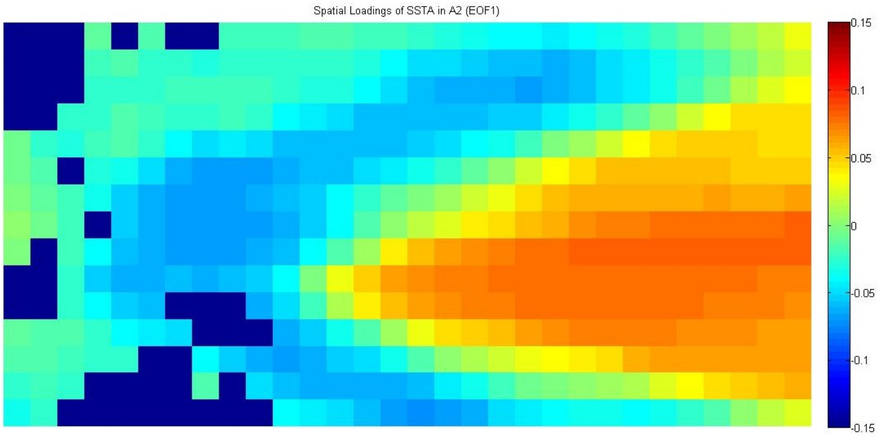

Figure 5.21: Spatial pattern of EOF1 in area A2. ENSO is the leading area in the system. Lands are displayed in blue, in particular Eastern China is displayed in the upper left part, North of Australia is displayed in bottom left (see figure 5.2 to have more spatial reference).

evaluated for each combination of SSTA areas and streamflow/rainfall anomalies (see Table 5.1.3). CCA is calculated with different time lag: from time lag 0 to time lag 6.

Number of EOF for hydrological variables (predictand)	Number of EOF for SSTA (predictor)
3 EOFs of rainfall anomalies	20 EOFs of SSTA in A1
3 EOFs of rainfall anomalies	29 EOFs of SSTA in A2
2 EOFs of rainfall anomalies	20 EOFs of SSTA in A1
2 EOFs of rainfall anomalies	29 EOFs of SSTA in A2

Table 5.2: Number of EOF selected in input to CCA.

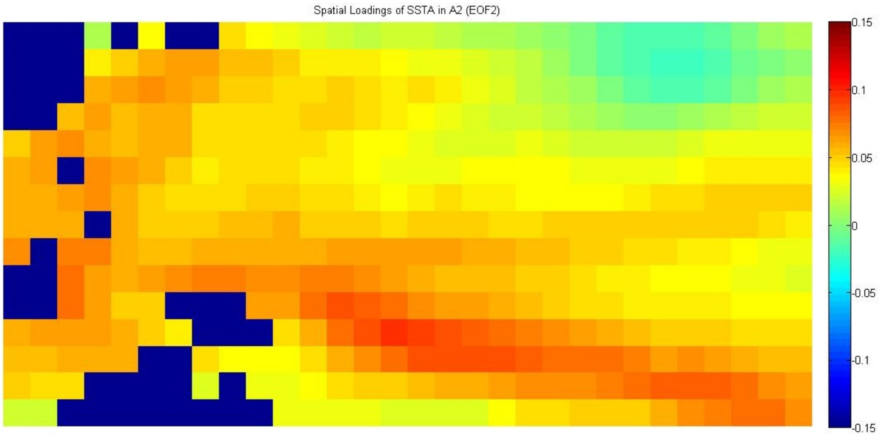

Figure 5.22: Spatial pattern of EOF2 in area A2. Lands are displayed in blue, in particular Eastern China is displayed in the upper left part, North of Australia is displayed in bottom left (see figure 5.2 to have more spatial reference).

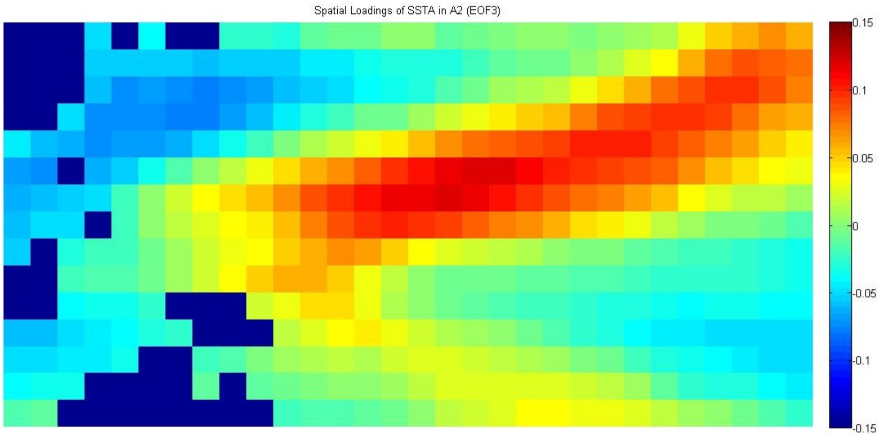

Figure 5.23: Spatial pattern of EOF3 in area A2. Lands are displayed in blue, in particular Eastern China is displayed in the upper left part, North of Australia is displayed in bottom left (see figure 5.2 to have more spatial reference).

Results of CCA show always a positive correlation among the first principal component evaluated in output of CCA for the predictor and the first principal components evaluated in output for the predictand. The values of the correlation coeffient are almost costant in each time lag for each combination of CCA. Bar graphs in figure 5.25 and in figure 5.26 display the correlation coefficients of rainfall experiments in area A1 and and in area A2 for each time lag. The correlation coefficients are around 30%, in particular, in area A1 it varies from 20% to 30% and in area A2 from 25% to 30%. Bar graphs of streamflow experiments are shown in figure 5.27 and figure 5.28 and correlation coefficients are around 40%: in area A1 it varies around from 35% to 40%, and in area A2 it is costant around 40%. As a consequence of

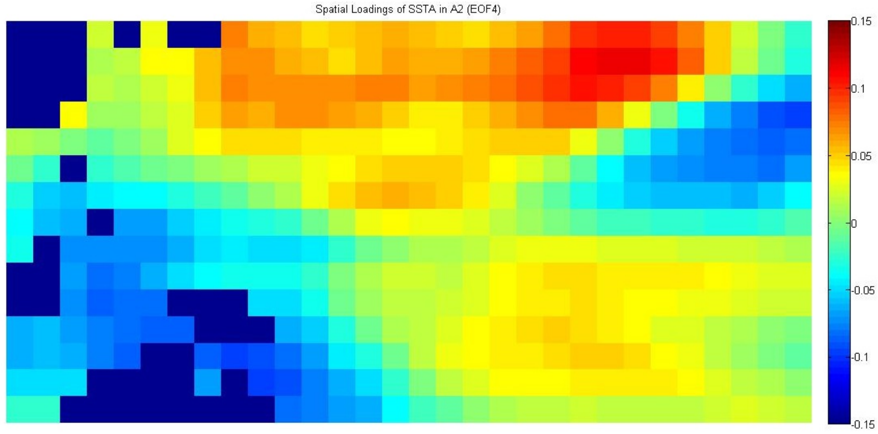

Figure 5.24: Spatial pattern of EOF4 in area A2. Lands are displayed in blue, in particular Eastern China is displayed in the upper left part, North of Australia is displayed in bottom left (see figure 5.2 to have more spatial reference).

these results, a existence of teleconnections between SSTA in the two ocean areas and the rainfall/streamflow anomalies is proven by CCA method.

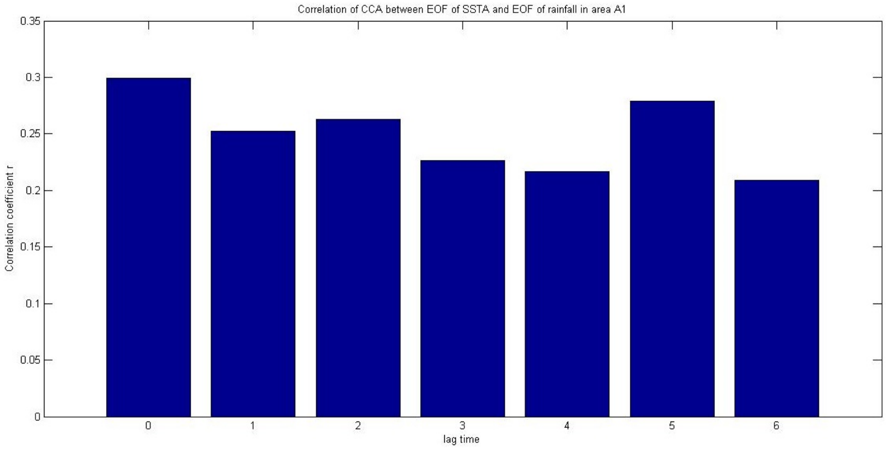

Figure 5.25: Correlation coefficient of CCA between EOF of rainfall anomalies and EOF of SSTA in area A1 at different time lag.

Similar patterns are found plotting the first CCA's principal components of EOF of SSTA and EOF of rainfall/streamflow anomalies in some months. An example is shown in figure 5.29, in which the first CCA's principal components of CCA between EOF of rainfall anomalies and EOF of SSTA in A1 at lag 1 are displayed around from August 1998 and October 2002. In June 1999 and December 2002 the trend of the two series are similar. Not in all months the two signal are in-phase, but it is justified by the corresponding correlation of the two series, that is around 30%. An example for streamflow anomalies

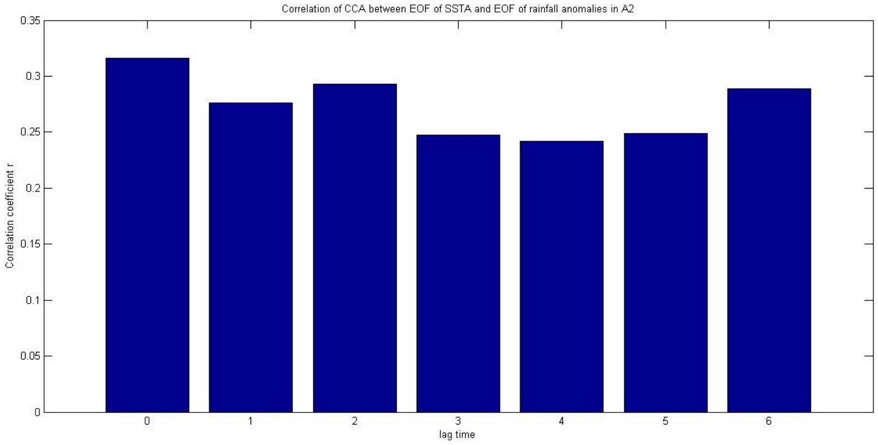

Figure 5.26: Correlation coefficient of CCA between EOF of rainfall anomalies and EOF of SSTA in area A2 at different time lag.

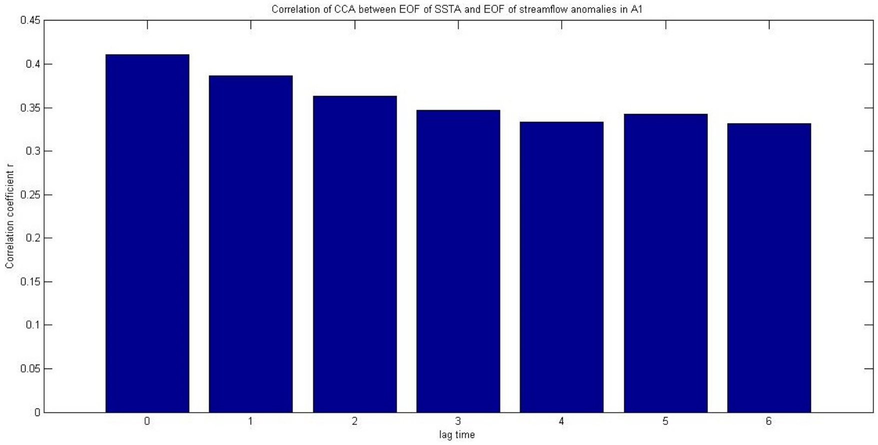

Figure 5.27: Correlation coefficient of CCA between EOF of streamflow anomalies and EOF of SSTA in area A1 at different time lag.

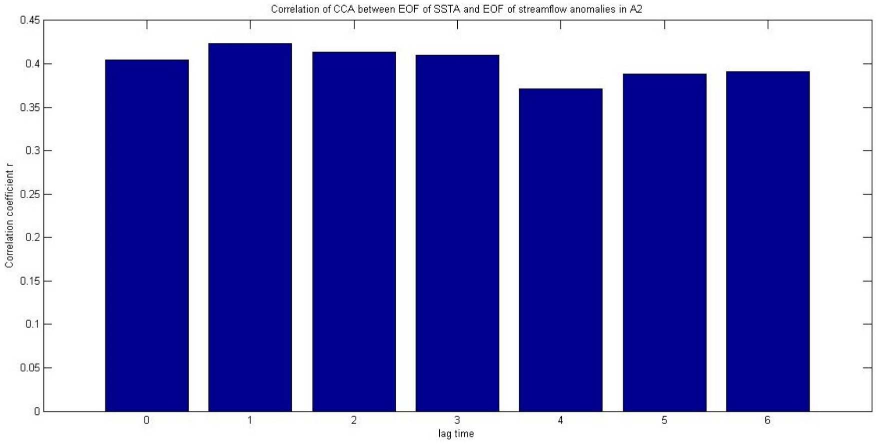

Figure 5.28: Correlation coefficient of CCA between EOF of streamflow anomalies and EOF of SSTA in area A2

is shown in figure figure 5.30, where the first CCA's principal components of streamflow anomalies and SSTA in area A2 at lag 1 are displayed. In particulat similar trend are found before August 1998 and before April 1980. As for rainfall anomalies, there is not a perfect in-phase signal because the correlation coefficients among the two series is around 40%.

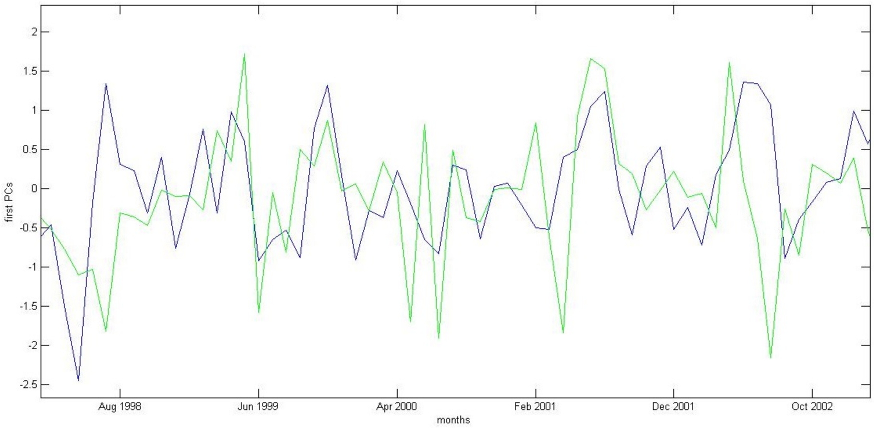

Figure 5.29: The first CCA's principal components of rainfall anomalies and SSTA in A1 at lag 1. The correlation coefficients between the two series is around 30%. Time period: August 1998 - October 2002.

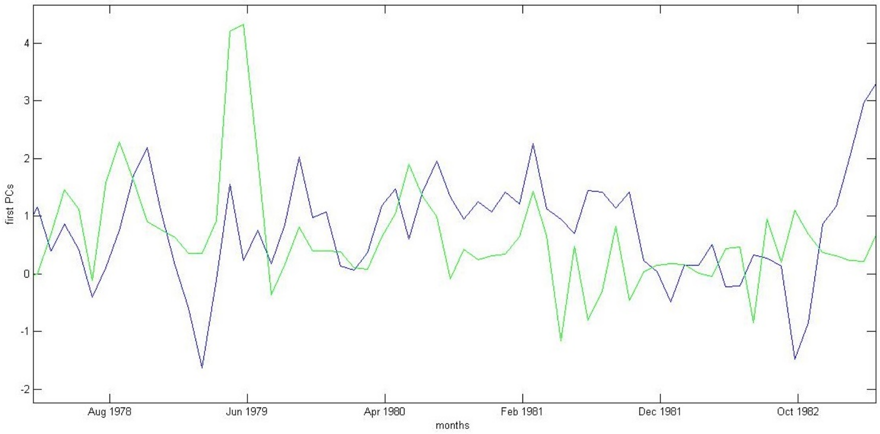

Figure 5.30: The first CCA's principal components of streamflow anomalies and SSTA in A2 at lag 1. The correlation coefficients between the two series is around 40%. Time period: August 1978 - October 1982.

After CCA analysis, a correlation analysis between the first CCA's principal components of the EOF of SSTA and SSTA at each grid point is computed. The aim of this kind of analysis is to evaluate if a particular part of SSTA is more correlated with the results of CCA analysis in order to understand if there is a part in the oceans that is greater linked with rainfall and streamflow anomalies. Interesting results are found especially when CCA is computed with EOF of streamflow anomalies. Results evaluated in area A2 show always a positive and significant correlation in the ENSO area. The correlation is widespread in ENSO area with the number of time lag (see figure 5.31). This results mean that the first CCA's principal components, that have a correlation coefficients of around 40% with EOF of streamflow anomalies, are strongly related with ENSO area. As a consequence, ENSO area is linked with streamflow anomalies and probably a teleconnection between SST in Pacific Ocean and streamflow exists. Results calculated in area A1 have not positive correlations if the correlation is computed with the first principal components from lag 0 to lag 2 (e.g. see figure 5.32), after lag time 3 until lag time 6 (see figure 5.33) the maps correlation show strong positive correlation, especially in the Indian Ocean and in China sea. These results are always significant. They show a delay in the correlations and, as a consequence, the SSTA in Indian Ocean and China sea have effects on streamflow after 3 months.

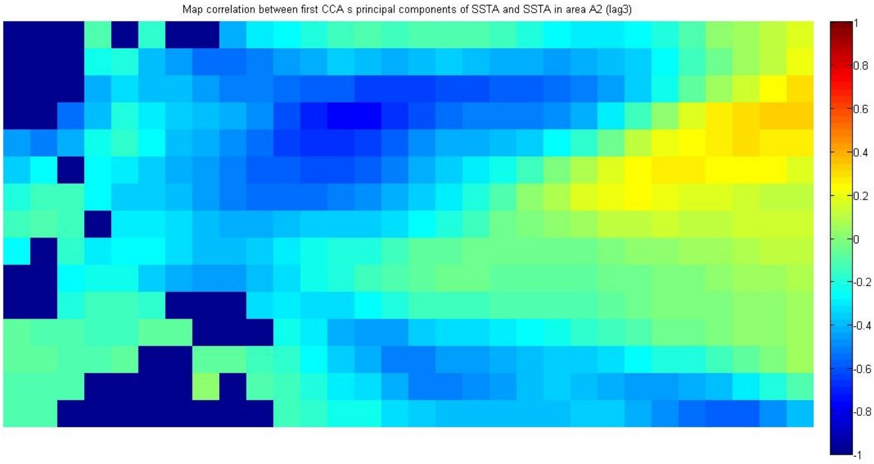
Map correlation between first CCA s principal components of SSTA and SSTA in area A2 (lag3)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 5.31: Map correlation between first CCA's principal components of SSTA and SSTA in area A2 at lag 3 in streamflow analysis. ENSO area is positive correlated. Other correlation maps show similar results. Lands are displayed in blue, in particular Eastern China is displayed in the upper left part, North of Australia is displayed in bottom left (see figure 5.2 to have more spatial references).

The same correlation analysis are evaluated also for rainfall but results do not show a regular pattern as for streamflow. Correlation analysis evaluated between first CCA's principal components of EOF of SSTA in area A2 and rainfall anomalies show positive correlation in ENSO area but the shape of correlation varies more with lag time instead as for streamflow (see figure 5.34). Moreover correlation analysis evaluated on area A1 show spots widespread over Indian Ocean, Indonesian sea and China sea that change at each lag time (see figure

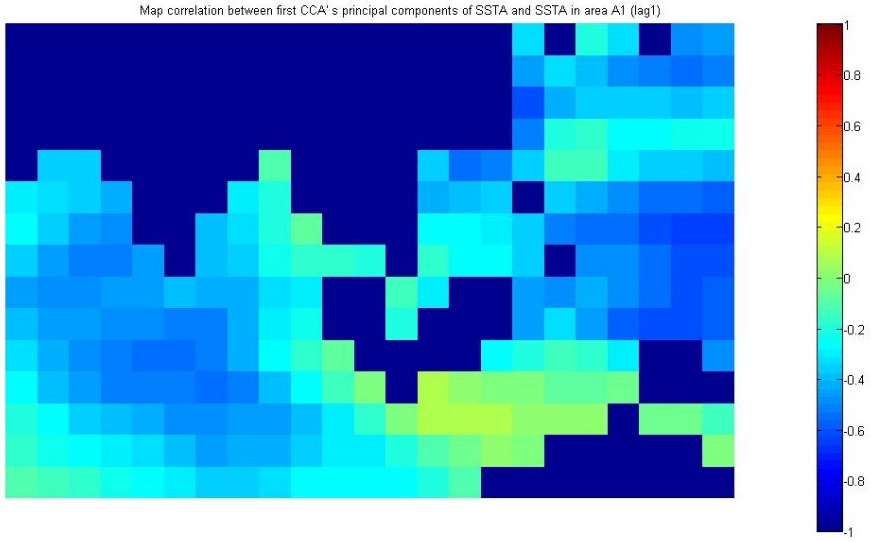
Map correlation between first CCA' s principal components of SSTA and SSTA in area A1 (lag1)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 5.32: Map correlation between first CCA's principal components of SSTA and SSTA in area A1 at lag 1 in streamflow analysis. No positive correlation is observed. Lands are displayed in blue, in particular Southeast Asia is displayed in the upper left part, North of Australia is displayed in bottom right (see figure 5.1 to have more spatial references).

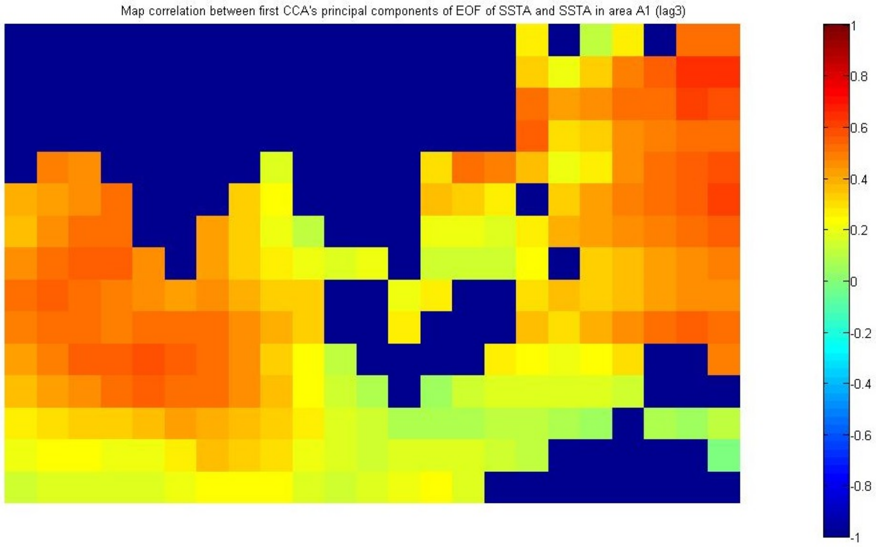
Map correlation between first CCA's principal components of EOF of SSTA and SSTA in area A1 (lag3)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 5.33: Map correlation between first CCA's principal components of SSTA and SSTA in area A1 at lag 3 in streamflow analysis. Positive correlation is observed in Indian Ocean and China sea. Lands are displayed in blue, in particular Southeast Asia is displayed in the upper left part, North of Australia is displayed in bottom right (see figure 5.1 to have more spatial references).

5.35).

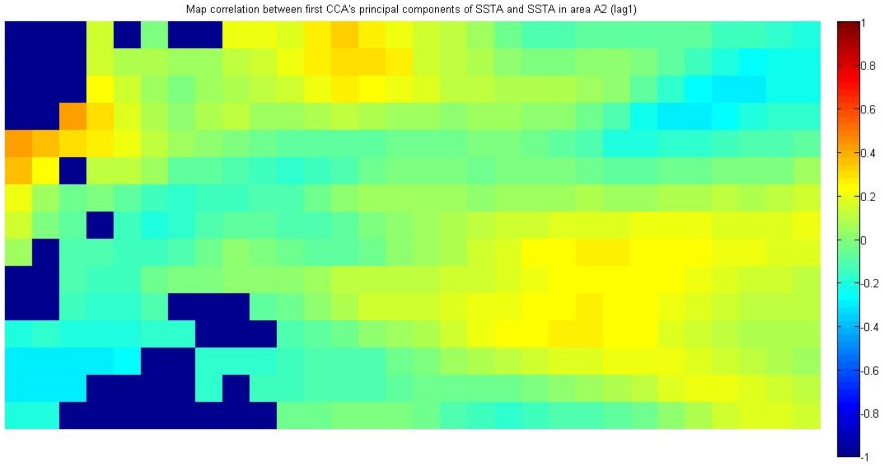
Map correlation between first CCA's principal components of SSTA and SSTA in area A2 (lag1)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 5.34: Map correlation between first CCA's principal components of SSTA and SSTA in area A2 at lag 1 in rainfall analysis. Lands are displayed in blue, in particular Eastern China is displayed in the upper left part, North of Australia is displayed in bottom left (see figure 5.2 to have more spatial references).

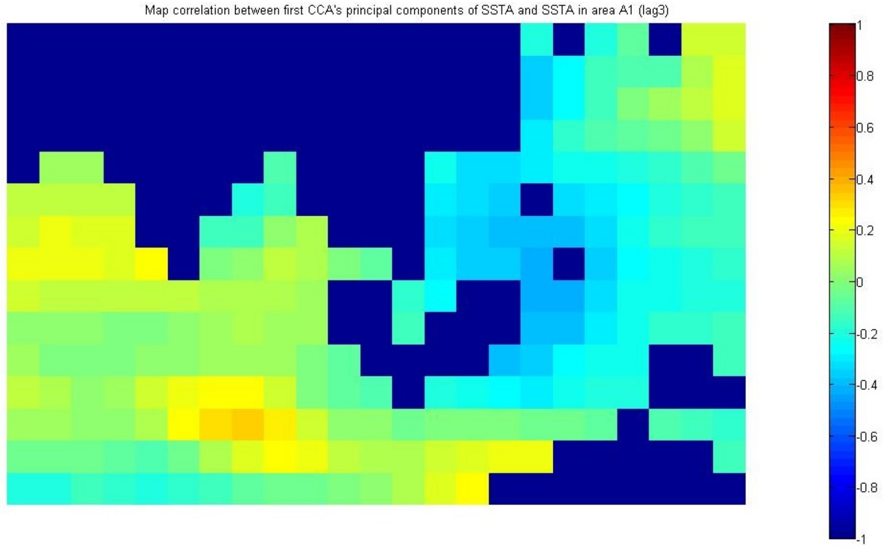
Map correlation between first CCA's principal components of SSTA and SSTA in area A1 (lag3)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 5.35: Map correlation between first CCA's principal components of SSTA and SSTA in area A1 at lag 3 in rainfall analysis. Lands are displayed in blue, in particular Southeast Asia is displayed in the upper left part, North of Australia is displayed in bottom right (see figure 5.1 to have more spatial references).

6 STREAMFLOW PREDICTION

6.1 introduction

The second part of this study is to build prediction models of the 11 station's streamflow using as input variables the EOF's principal compontents of SSTAs in area A1 and A2 got from the previous analysis. The aim is to evaluate the performance of the models, assuming that a teleconnection exists between streamflow and SSTAs in Pacific Ocean and Indian Ocean as it is discussed in chapter 6. Two methods are used to achieve it: CCA and IVS. Similarly to CCA's analysis, the number of principal components retained for each area is chosen following the Kaiser method, such as 20 principal components in area A1 and 29 principal components in area A2. The coefficient of determination R^2 is the index used to evaluate the performance.

6.2 cca's prediction

CCA can be used to make prediction using the CCA's amplitude time series of predictor (or predictand) (see section 3.3.3). CCA's prediction models are built evaluating CCA between the streamflow anomalies recorded in the 11 stations of the basin and the SSTA's principal components in each area (A1 or A2). The models are evaluated separately for each area: SSTA's principal components in area A1 evaluate a model that concerns only SSTAs in area A1, and the SSTA's principal components in area A2 for the model related to area A2. The models are calculated for different time lags between the streamflow anomalies and the SSTA's principal components, from lag of 0 months to a lag of 6 months. The outputs of CCA's prediction model are predicted streamflow anomalies series for each station (11 time series) thus R^2 indexes are calculated to evaluate the performance in every station. R^2 are evaluated by a k-fold cross-validation with 10 number of folds.

R^2 indexes in calibration relative models evaluated using SSTA of area A1, are summarized in tab.6.1 and tab.6.2. The best performances does not exceed 0.19 of R^2 and is found in the first lag time. In general, indexes R^2 in area A1 decrease with lags, but ΔR^2 varies little from a lag time to another (the maximum variation of R^2 is around 0.02) but there are some exceptions. For example in ChiemHoa, LaiChau and ThuongCat R^2 in lag of 2 months is lower than R^2 in lag of 3 months, or in HaNoi, SonTay, VuQuang and ThuongCat R^2 in lag of 5 months is lower than R^2 in lag of 6 months.

Performances concerning area A2 are similar to those evaluated in relation to area A1, the difference is that the decrease with lag is not evident here as in the previous model. R^2 values are reported in tab.6.3 and tab.6.4. R^2 below 0.05 is not found in spite of in model evaluated at area A2 reaches R^2 of 0.02.

In summary, the R^2 values in calibration in the CCA's models have the highest value around 0.20 in NamMuc station. In general R^2 higher values in each station vary from 0.10 to 0.15, and they depend on the particular station.

lag	ChiemHoa	HamYen	HaNoi	HoaBinh	LaiChau	NamMuc
0	0.1104	0.1100	0.1396	0.1466	0.1145	0.1900
1	0.0848	0.0754	0.1039	0.1054	0.0745	0.1635
2	0.0723	0.0555	0.0983	0.0979	0.0686	0.1494
3	0.0763	0.0676	0.0849	0.0894	0.0706	0.1439
4	0.0428	0.0443	0.0637	0.0745	0.0629	0.1353
5	0.0373	0.0276	0.0474	0.0685	0.0729	0.1301
6	0.0422	0.0249	0.0567	0.0636	0.0624	0.1238

Table 6.1: R^2 in calibration in CCA's prediction models of streamflow anomalies in area A1-part1. Station from ChiemHoa to NamMuc (see figure 4.5).

lag	SonTay	TaBu	ThuongCat	VuQuang	YenBai
0	0.1313	0.1512	0.1441	0.1100	0.1626
1	0.0962	0.0986	0.1121	0.0864	0.1208
2	0.0871	0.0908	0.1154	0.0700	0.1222
3	0.0858	0.0876	0.1082	0.0765	0.1165
4	0.0782	0.0720	0.0881	0.0476	0.0784
5	0.0614	0.0779	0.0665	0.0359	0.0617
6	0.0471	0.0687	0.0603	0.0465	0.0545

Table 6.2: R^2 in calibration in CCA's prediction models of streamflow anomalies in area A1-part2. Station from SonTay to YenBai (see figure 4.5).

Performances in validation are always worse than in calibration. Because R^2 values in calibration don't show a strong performance, high values of R^2 are not expected in validation. They vary from -0.20 to 0 and a positive value is never observed. Below are reported plots of streamflow anomalies in NamMuc station evaluated at time lag 1 in calibration (see figure 6.1 for area A1 and figure 6.2 for area A2) and

lag	ChiemHoa	HamYen	HaNoi	HoaBinh	LaiChau	NamMuc
0	0.1091	0.1028	0.1263	0.1164	0.1082	0.2025
1	0.0646	0.0726	0.0869	0.1003	0.0943	0.1613
2	0.0692	0.0608	0.1070	0.1131	0.1032	0.1690
3	0.0890	0.0876	0.1166	0.1215	0.1117	0.2079
4	0.1021	0.0922	0.1059	0.0782	0.0750	0.1890
5	0.0813	0.0727	0.0857	0.0642	0.0564	0.1669
6	0.0785	0.0780	0.0965	0.0624	0.0614	0.1526

Table 6.3: R^2 in calibration in CCA's prediction models of streamflow anomalies in area A2-part1. Stations from ChiemHoa to NamMuc (see figure 4.5).

lag	SonTay	TaBu	ThuongCat	VuQuang	YenBai
0	0.1100	0.1322	0.1580	0.1222	0.1567
1	0.0710	0.1055	0.1066	0.0715	0.1066
2	0.0769	0.1103	0.1226	0.0848	0.1209
3	0.0978	0.1256	0.1359	0.1216	0.1255
4	0.0969	0.0872	0.1083	0.1306	0.1050
5	0.0809	0.0655	0.0812	0.1078	0.0990
6	0.0889	0.0714	0.0789	0.1194	0.1052

Table 6.4: R^2 in calibration in CCA's prediction models of streamflow anomalies in area A2-part2. Stations from SonTay to YenBai (see figure 4.5).

in validation (see figure 6.3 for area A1 and figure 6.4 for area A2).

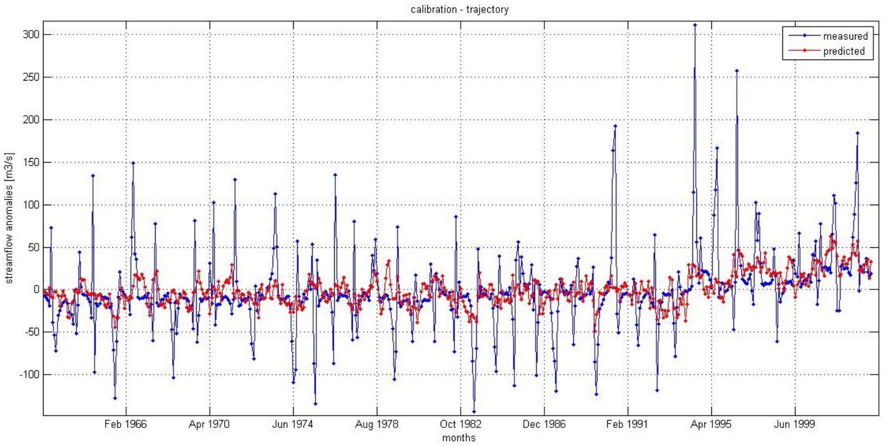

Figure 6.1: Plot CCA's prediction model in calibration in area A1 time lag 1. Station: NamMuc.

The previous models are evaluated on the streamflow series obtained by adding the monthly average to the streamflow anomalies predicted by CCA. These analysis are computed to test the contribution

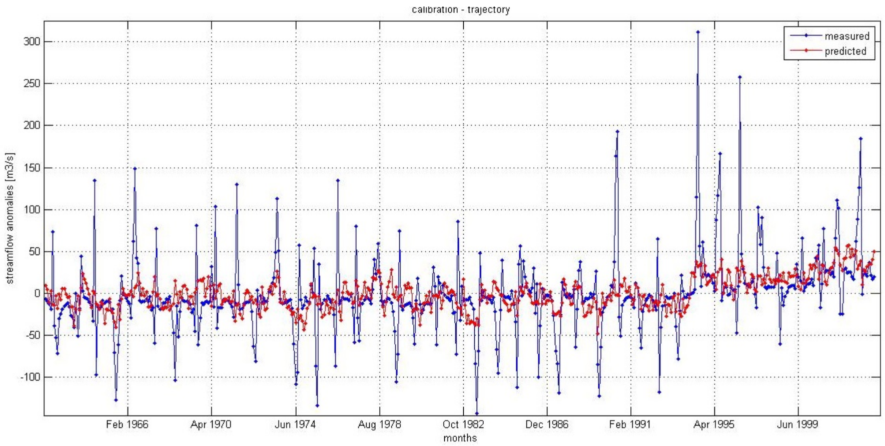

Figure 6.2: Plot CCA's prediction model in calibration in area A2 time lag 1. Station: NamMuc.

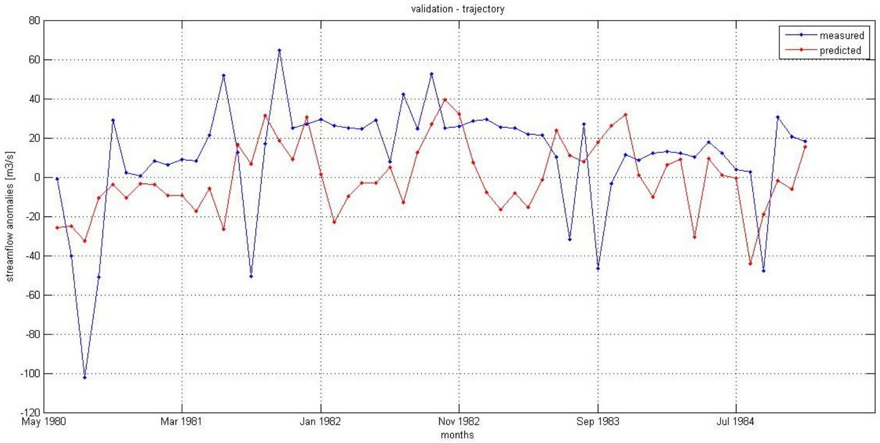

Figure 6.3: Plot CCA's prediction model in validation in area A1 time lag 1. Station: NamMuc.

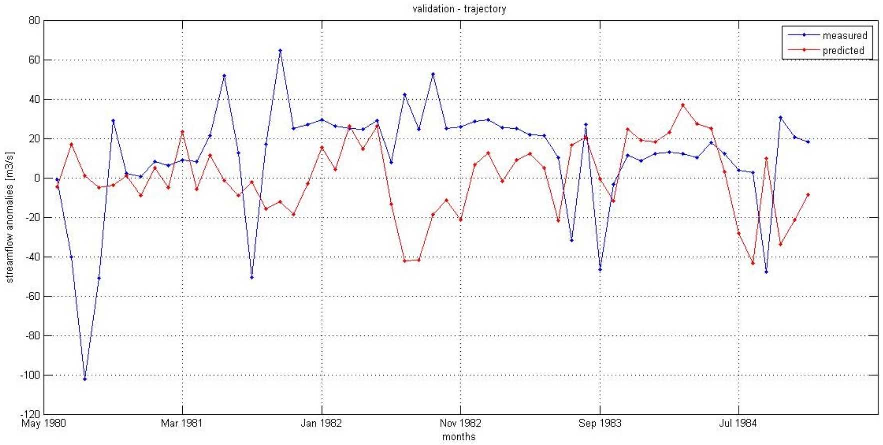

Figure 6.4: Plot CCA's prediction model in validation in area A2 time lag 1. Station: NamMuc.

of the monthly average in the forecasting. The R^2 values are higher than those evaluted for streamflow anomalies because the monthly average improves the performances. The calibration values of R^2 are lower than validation values of R^2. Better performances are found in validation because the streamflow anomalies predicted previously have bad performances thus they have a low contribution in the model. In other words, the validation models are like models of the only monthly averages related to the streamflow series. R^2 values in calibration are reported in tab.6.5, tab.6.6, tab.6.9 and tab.6.10. R^2 values in validation are reported in tab.6.7, tab.6.8, tab.6.11 and tab.6.12.

lag	ChiemHoa	HamYen	HaNoi	HoaBinh	LaiChau	NamMuc
0	0.6712	0.7397	0.7671	0.7625	0.7852	0.6694
1	0.6779	0.7470	0.7765	0.7752	0.7979	0.6879
2	0.6830	0.7533	0.7772	0.7747	0.7977	0.6906
3	0.6826	0.7510	0.7795	0.7763	0.7957	0.6911
4	0.6869	0.7531	0.7809	0.7762	0.7966	0.6845
5	0.6899	0.7596	0.7843	0.7800	0.7999	0.6896
6	0.6887	0.7600	0.7828	0.7828	0.8016	0.6869

Table 6.5: R^2 in calibration in CCA's prediction models of streamflow in area A1-part1. Station from ChiemHoa to NamMuc (see figure 4.5).

lag	SonTay	TaBu	ThuongCat	VuQuang	YenBai
0	0.7634	0.7726	0.7481	0.7085	0.6617
1	0.7732	0.7899	0.7561	0.7137	0, 6751
2	0.7751	0.7906	0.7520	0.7200	0.6784
3	0.7744	0.7889	0.7559	0.7192	0.6761
4	0.7735	0.7892	0.7567	0.7211	0.6834
5	0.7748	0.7931	0.7635	0.7258	0.6867
6	0.7784	0.7944	0.7672	0.7221	0.6905

Table 6.6: R^2 in calibration in CCA's prediction models of streamflow in area A1-part2. Station from SonTay to YenBai (see figure 4.5).

6.3 IVS'S PREDICTION

IVS method is used in two steps: in the first step variables, which best represent the relationship with streamflow anomalies, are ranked; in the second step, a multi-step prediction model is developed using the selected variables as input to forecast streamflows anomalies. IIS algorithm is used to develop IVS, with an ensemble of 250 trees and a predefined threshold n_{min} of 5. The performance is evaluated by a 5 fold cross-validation. The aim is to forecast streamflow anomalies having EOF's principal components of different lead time. Three IIS

lag	ChiemHoa	HamYen	HaNoi	HoaBinh	LaiChau	NamMuc
0	0.7205	0.8785	0.7905	0.7928	0.8286	0.7271
1	0.7201	0.8853	0.7937	0.8001	0.8261	0.7466
2	0.7395	0.9096	0.8173	0.8113	0.8310	0.7608
3	0.7320	0.9012	0.8077	0.8099	0.8328	0.7776
4	0.7245	0.9001	0.8030	0.7976	0, 8313	0.8402
5	0.7202	0.9044	0.8201	0.8181	0.8286	0.8514
6	0.7055	0.9004	0.8017	0.8107	0.8239	0.8426

Table 6.7: R^2 in validation in CCA's prediction models of streamflow in area A1-part1. Station from ChiemHoa to NamMuc (see figure 4.5).

lag	SonTay	TaBu	ThuongCat	VuQuang	YenBai
0	0.7615	0.8613	0.8412	0, 7732	0.7842
1	0.7621	0.8570	0.8459	0.7707	0.7901
2	0.7851	0.8650	0.8504	0.8073	0.7995
3	0.7801	0.8659	0.8527	0.7986	0.7921
4	0.7659	0.8677	0.8391	0.7805	0.8172
5	0, 8056	0.8720	0.8600	0.7862	0.8357
6	0.7914	0.8646	0.8498	0.7708	0.8155

Table 6.8: R^2 in validation in CCA's prediction models of streamflow in area A1-part2. Station from SonTay to YenBai (see figure 4.5).

lag	ChiemHoa	HamYen	HaNoi	HoaBinh	LaiChau	NamMuc
0	0.6729	0.7437	0.7696	0.7718	0.7906	0.6709
1	0.6843	0.7523	0.7817	0.7753	0.7945	0.6864
2	0.6801	0.7522	0.7757	0.7721	0.7925	0.6801
3	0.6799	0.7448	0.7715	0.7656	0.7857	0.6731
4	0.6717	0.7449	0.7738	0.7786	0.7959	0.6772
5	0.6813	0.7524	0.7776	0.7820	0.8011	0.6832
6	0.6705	0.7470	0.7721	0.7796	0.7974	0.6811

Table 6.9: R^2 in calibration in CCA's prediction models of streamflow in area A2-part1. Station from ChiemHoa to NamMuc (see figure 4.5).

experiments are computed to assess the relationship between EOF's principal components of SSTA and streamflow anomalies in a station, at each experiment different time lags are evaluated. Experiments are evaluated dividing EOF's principal components for each area (A1 and A2). The three experiments are summarized below, where "EOF" means a set of EOF's principal components. A set is the number of principal components chosen for each kind of model (summarized in tab.6.13). These numbers are selected observing the R^2 of each linear model solved by least square method. In particular the numbers of

lag	SonTay	TaBu	ThuongCat	VuQuang	YenBai
0	0.7672	0.7822	0.7451	0.7065	0.6623
1	0.7798	0.7872	0.7568	0.7196	0.6796
2	0.7760	0.7861	0.7516	0.7145	0.6777
3	0.7707	0.7780	0.7485	0.7056	0.6706
4	0.7700	0.7892	0.7551	0.7017	0.6757
5	0.7729	0.7936	0.7630	0.7119	0.6779
6	0.7679	0.7894	0.7592	0.6976	0.6725

Table 6.10: R^2 in calibration in CCA's prediction models of streamflow in area A2-part2. Station from SonTay to YenBai (see figure 4.5).

lag	ChiemHoa	HamYen	HaNoi	HoaBinh	LaiChau	NamMuc
0	0.7307	0.8835	0.7861	0.7923	0.8242	0.7267
1	0.7458	0.8886	0.8091	0.8047	0.8310	0.7554
2	0.7339	0.9028	0.8060	0.7984	0.8222	0.7685
3	0.7136	0.8979	0.8050	0.7949	0.8236	0.7277
4	0.6914	0.8857	0.7880	0.7925	0.8302	0.8106
5	0.6791	0.8797	0.7938	0.8167	0.8320	0.8233
6	0.6997	0.8748	0.7827	0.8095	0.8276	0.8289

Table 6.11: R^2 in validation in CCA's prediction models of streamflow in area A2-part1. Station from ChiemHoa to NamMuc (see figure 4.5).

lag	SonTay	TaBu	ThuongCat	VuQuang	YenBai
0	0.7637	0.8561	0.8390	0.7622	0.7821
1	0.7874	0.8603	0.8574	0.7884	0.8017
2	0.7861	0.8574	0.8367	0.7960	0.7949
3	0.7769	0.8538	0.8308	0.7825	0.7897
4	0.7555	0.8635	0.8335	0.7419	0.8071
5	0.7839	0.8688	0.8456	0.7273	0.8027
6	0.7690	0.8659	0.8438	0.7355	0.7950

Table 6.12: R^2 in validation in CCA's prediction models of streamflow in area A2-part2. Station from SonTay to YenBai (see figure 4.5).

principal components are selected when the linear model have an R^2 of at least 0.1.

$$\text{model 1} \quad y_t = EOF(SSTA(t-1)) + EOF(SSTA(t-2)) + EOF(SSTA(t-3))$$

$$\text{model 2} \quad y_t = EOF(SSTA(t-2)) + EOF(SSTA(t-3)) + EOF(SSTA(t-4))$$

$$\text{model 3} \quad y_t = EOF(SSTA(t-3)) + EOF(SSTA(t-4)) + EOF(SSTA(t-5))$$

The previous models are evaluated also with in input the streamflow anomalies at time lags: t-1 in model 1, t-2 in model 2 and t-3 in model

Model	A1 area	A2 area
model 1	13	15
model 2	14	18
model 3	19	19

Table 6.13: Number of principal components evaluated in each model

3.

IIS Prediction Model 1
Experiments of model 1 select the first principal compontents from each ocean area (A1 and A2) at first positions. In general rankings are similar in every station, this evidence proves that a comparable relationships exist between streamflow anomalies and principal components over the basin. For model evaluated in area A1, the second principal components at time lag 1 is the most selected. In particular it is ranked at first position by IIS in 10 of the 11 stations, in particular 7 of the 10 stations select it more than 5 times in 10 total times (see tab.6.14). This result is noticeable thinking that IIS evaluates the importance of 39 total inputs. Observing the second principal component of area A1, IIS selects the ocean areas above Australia and close to Philippines that are so the most affecting streamflow anomalies. IIS results concerning area A2 select the first four principal components at time lag 1, especially the fourth principal component. For example in YenBai station, the fourth principal component is ranked in first position 8 times over 10 times, and the first principal component is ranked in the second one 5 times over 10 times. In Sontay the fourth principal component is ranked at first position 6 times over 10 times. In ThuongCat, the fourth principal component and the first principal component are ranked in first position 4 and 2 times, principal components that appear also in second position 2 and 1 times. Both these results suggest that ENSO area affect streamflow anomalies, infact ENSO area is the most prominent cluster of cells in the first principal component and one of the leading part of ocean on the fourth principal component.

Model performances in area A1 with only the second component as inpunt are all around 0.50 of R^2 in calibration and -0.10 of R^2 in validation. In general results of IIS are better than those observed in CCA. For example performances are summarized in tab.6.15. In tab.6.16 performances using in input also streamflow anomalies at t-1 are shown. Performances of model in area A2 using the first and the fourth principal components are reported in tab.6.17, and in tab.6.18 using also the streamflow anomalies at t-1 as input. In general there is not high results of R^2 index in validation. The highest value of R^2 in validation

Station	Freq. 1° position	Freq. 2° position	Freq. 3° position
ChiemHoa	5	2	3
HamYen	1	1	2
HaNoi	7	0	1
HoaBinh	7	0	1
LaiChau	2	0	1
NamMuc	0	0	0
SonTay	3	0	1
TaBu	7	1	0
ThuongCat	8	0	0
VuQuang	6	0	1
YenBai	5	0	0

Table 6.14: Frequency of the second principal component in area A1 at time t-1. The frequency refers to a 10 total times of run.

is around 0.08 and it is reached using as input the streamflow anomalies at t-1.

Station	1 iteration	2 iteration
Input of model	2° PC (t-1)	1° PC(t-1)
TaBu		
calibration	0.5690	0.6901
validation	−0.1044	−0.0700
VuQuang		
calibration	0.5763	0.7225
validation	−0.0971	−0.0835

Table 6.15: R^2 in calibration and validation of model 1 in A1. Input variable are the second principal component and the first principal component. Often only the second principal component is the only ranked.

IIS Prediction Model 2

Models 2 do not show interesting results thus not a particular selection of principal components is selected by IIS. It is only noticeable that the first four principal components at time t-3 and t-4 are the most selected in area A1. In particular the third principal component at time t-4 is the most selected, and models with only this variable as input have around 0.58 of R^2 in calibration and around −0.10 of R^2 in validation. In area A2 the first four principal components at time t-2, t-3 and t-4 are the most ranked, without a characteristic pattern in the selection of input. Specifically IIS selects the first and the fourth principal components at each time lag, in particular the fourth principal component is only selected at time t-2 and t-3. As

Station	1 iteration	2 iteration	3 iteration
Input of model	flow anom. (t-1)	3° PC (t-1)	2° PC (t-1)
HaNoi			
calibration	0.6186	0.7624	0.7895
validation	−0.0416	0.0500	0.0828
VuQuang			
calibration	0.6198	0.7573	0.7916
validation	−0.0568	0.0161	0.0867

Table 6.16: R^2 in calibration and validation of model 1 in A1 with streamflow anomalies in input. Input are the streamflow anomalies at t-1, the third principal component and the second principal component at t-1.

Station	1 iteration	2 iteration
Input of model	1° PC (t-1)	2° PC (t-1)
YenBai		
calibration	0.5899	0.7291
validation	−0.1309	−0.0407
SonTay		
calibration	0.5697	0.7243
validation	−0.1309	−0.0560

Table 6.17: R^2 in calibration and validation of model 1 in A2. The input of the model are in order of iteration: the first principal component and the second principal component at time t-1.

Station	1 iteration	2 iteration
Input of model	flow anom. (t-1)	3° PC (t-1)
HoaBinh		
calibration	0.6022	0.7340
validation	−0.0872	−0.0481
Input of model	flow anom. (t-1)	1° PC (t-1)
SonTay		
calibration	0.5852	0.7515
validation	−0.0803	0.0806

Table 6.18: R^2 in calibration and validation of model 1 in A2 with streamflow anomalies in input.

a consequence, ENSO area is once again proved to affect the streamflow anomalies. Performances of models in area A1 are reported in tab.6.19 and tab.6.20, performances of models in area A2 are reported in tab.6.21 and tab.6.22.

Station	1 iteration	2 iteration	3 iteration
Input of model	1° PC (t-3)	2° PC (t-2)	3° PC (t-4)
TaBu calibration validation	 0.5308 −0.1550	 0.6970 −0.1346	 0.7229 −0.0652
Input of model	1° PC (t-3)	3° PC (t-4)	1° PC (t-2)
TaBu calibration validation	 0.5713 −0.1271	 0.7193 −0.0126	 0.7440 −0.0110

Table 6.19: R^2 in calibration and validation of model 2 in A1.

Station	1 iteration	2 iteration	3 iteration
Input of model	flow anom. (t-2)	2° PC (t-4)	1° PC (t-3)
NamMuc calibration validation	 0.5449 −0.1037	 0.7068 −0.0352	 0.7633 0.0446
Input of model	flow anom. (t-2)	2° PC (t-2)	
ThuongCat calibration validation	 0.5755 −0.0773	 0.7192 −0.0722	

Table 6.20: R^2 in calibration and validation of model 2 in A1 with streamflow anomalies in input.

Station	1 iteration	2 iteration
Input of model	1° PC (t-4)	4° PC (t-2)
VuQuang calibration validation	 0.5407 −0.1566	 0.7356 −0.0338
Input of model	2° PC (t-2)	4° PC (t-3)
NamMuc calibration validation	 0.5753 −0.1086	 0.7212 0.0327

Table 6.21: R^2 in calibration and validation of model 2 in A2.

IIS Prediction Model 3

A particular pattern of selection of input variables for every station is not found also in models 3. As in the previous models, often the first four principal components are selected but without a regular pattern. In general, the most selected principal components in area A1 is the third principal component at time t-4 that is the same input selected

Station	1 iteration	2 iteration
Input of model	flow anom. (t-2)	2° PC (t-4)
ThuongCat		
calibration	0.5780	0.7211
validation	−0.0968	−0.0646
Input of model	flow anom. (t-2)	1° PC (t-3)
YenBai		
calibration	0.6167	0.7233
validation	−0.0745	−0.0430

Table 6.22: R^2 in calibration and validation of model 2 in A2 with streamflow anomalies in input.

in model 2. This result confirm the role of the Indonesia sea that is the leading part in the third EOF's principal components in area A1. It affects the streamflow anomalies with a lag of 3 months. Results in area A2 are similar to those of model 2: a frequency in ranking of the first and the fourth principal components is noticed. Especially the fourth principal components are selected only in time t-3 and t-4, as in model 2 its contribution for long time lag is not important. Streamflow anomalies at time t-3 is not selected by IIS in most of the stations. As in the previous models, significant results are shown in tab.6.23 for area A1, and tab.6.24 for area A2.

Station	1 iteration	2 iteration
Input of model	3° PC (t-4)	1° PC (t-3)
HoaBinh		
calibration	0.5457	0.6756
validation	−0.1052	−0.0651
Input of model	1° PC (t-3)	3° PC (t-4)
LaiChau		
calibration	0.5470	0.6804
validation	−0.0816	−0.0804

Table 6.23: R^2 in calibration and validation of model 3 in A1.

In conclusion the only significant performances are found in calibration. Not good performances in validation using different selections of the first four principal components are found both for area A1 and area A2. In general, calibration performances are better than validation performances. Infact in validation models are tested with parameters evaluated with the calibration data set and this reduce the accuracy of the performance. In this work a k-fold cross-validation is used both for CCA and IIS, respectively 5 fold for CCA and 10 fold for IIS. The number of observations used to perform validation is lower than calibration, and that also affects the performances.

Station	1 iteration	2 iteration
Input of model	4° PC (t-3)	3° PC (t-3)
HoaBinh		
calibration	0.5544	0.7051
validation	−0.1195	−0.1045
Input of model	1° PC (t-3)	4° PC (t-3)
NamMuc		
calibration	0.5873	0.7296
validation	−0.0719	0.0131

Table 6.24: R^2 in calibration and validation of model 2 in A2.

7 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

The present thesis investigates the relationship between Indian and Pacific SST and hydro-meteorological variables, such as rainfall and streamflow, in the Red River Basin in Vietnam. With the purpose of improving the knowledge of the hydro-meteorological system in a basin that is undergoing rapid development in terms of population and economic growth. Therefore, studying the influence of SST on this fast-evolving context could be exploited for substantially improving water resources management with related significant socio-economic benefits.

Three main goals are considered and are summarized as follows: first to assess that effectively SST affects the hydro-meteorological system in the basin, second to select new indicator of SST in the Indian and Pacific ocean that best explain the previous relationship and last using the new SST's indicators in order to develop streamflow prediction models.
The most notable contribution of this work is the search for new kind of indicators of SST phenomena that can be related to teleconnection effects in a basin scale. In conclusion, EOF principal components of SSTA analysis can be used as indicators.

An EOF is developped on SSTA in order to compress information and capture the main process dynamics. The results of EOF are found to be consistent with the leading modes in Indian and Pacific Ocean. Especially the ENSO area is found to be the prominent one in the Pacific Ocean. A selection of the most significant EOF modes arrange a new set of SST's indicators. EOF are developped on streamflow anomalies and rainfall anomalies. EOFs of rainfall anomalies show a southward gradient that could be explained by the "corridor-barrier" phenomenon caused by the topography of the basin. CCA analysis, between the EOF results on rainfall/streamflow anomalies and EOF results of SST, are developped to assess the relationship between the SST's indicators and the hydro-meteorological variables over the basin. Observing the correlations between the output of CCA analysis, indicators of SST and the EOF of rainfall (streamflow) variables are proven to be related, thus around 30% (40%) of correlation coefficients are found. Finally streamflow prediction models are evaluated using in input the SST's indicators. IVS and CCA methods are computed to achieve it. CCA's models show low performances in calibration (around

0.10 of R^2) and zero performances in validation. IVS selects the most significant indicators as the most related to streamflow anomalies. However IVS's models show high performances in calibration but still zero performances in validation.

Further enhancements to this work would include the following aspect:

- In this work the effects of Indian and Pacific oceans are considered separately. Further researches may assess the effects of the two oceans studying a whole combined area. EOF results in this area should compare the processes in the two areas.
- ENSO-related effects may be compared with the indicator of Pacific SST, because some correspondences with ENSO phenomena are found in EOFs of this area. For example, some experiments of IVS could be evaluated in order to compare the importance of the SST indicator in streamflow forecasts.
- Predictions model could be improved by using in addition other meteorological variables (e.g. temperature, evaporation).

BIBLIOGRAPHY

[1] K. N. Amarasekera et al. "ENSO and the natural variability in the flow of tropical rivers." In: *Journal of Hydrology* 200.1-4 (1997), pp. 24–39 (cit. on pp. 2, 11).

[2] A. G. Barnston and T. M. Smith. "Specification and Prediction of Global Surface Temperature and Precipitation from Global SST using CCA." In: *Journal of Climate* 9 (1996), pp. 2660–2697 (cit. on pp. 1, 11, 13, 24).

[3] L. Beltrame and D. Carbonin. "ENSO teleconnection patterns on large scale water resources systems." Master Thesis. Politecnico di Milano, Milan, Italy, 2013 (cit. on p. 38).

[4] L. Breiman et al. *Classification and regression trees*. Wadsworth International Group, 1984. isbn: 0412048418 (cit. on p. 30).

[5] J. Chandimala and L. Zubair. "Predictability of stream flow and rainfall based on ENSO for water resources management in Sri Lanka." In: *Journal of Hydrology* 335.3-4 (2007), pp. 303–312 (cit. on p. 11).

[6] F. H. S. Chiew and T. A. McMahon. "Global ENSO- streamflow teleconnection, streamflow forecasting and interannual variability." In: *Hydrological Sciences Journal* 47.3 (2002), pp. 505–522 (cit. on pp. 1, 13).

[7] C. Deser et al. "Sea Surface Temperature Variability: Patterns and Mechanisms." In: *Annual Review of Marine Science* 2 (2010), pp. 115–143 (cit. on pp. 9, 15).

[8] S. Galelli and A. Castelletti. "Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling." In: *Hydrology and Earth System Sciences* 17 (2013a), pp. 2669–2684.

[9] S. Galelli and A. Castelletti. "Tree-based iterative input variable selection for hydrological modeling." In: *Water Resources Research* 49.7 (2013b), pp. 4295–4310 (cit. on pp. 27–29).

[10] S. Galelli et al. "An evaluation framework for input variable selection algorithms for environmental data-driven models." In: *Environmental Modelling Software* 62 (2014), pp. 33–51 (cit. on p. 26).

[11] P. Geurts, D. Ernst, and L. Wehenkel. "Extremely randomized trees." In: *Machine Learning* 63.1 (2006), pp. 3–42 (cit. on p. 30).

[12] K. Grantz and B. Rajagopalan. “A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts.” In: *Water Resources Research* 41 (2005) (cit. on p. 2).

[13] F. Gutierrez and J. A. Dracup. “An analysis of the feasibility of long-range streamflow forecasting for Colombia using El Niño Southern Oscillation indicators.” In: *Journal of Hydrology* 246.1-4 (2001), pp. 181–196 (cit. on p. 2).

[14] I. Guyon and A. Elisseeff. “An introduction to variable and feature selection.” In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182 (cit. on pp. 26, 27).

[15] M. I. Hejazi and X. Cai. “Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm.” In: *Advances in Water Resources* 32.4 (2009), pp. 582–593 (cit. on p. 26).

[16] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002. ISBN: 978-0-387-95442-4 (cit. on pp. 20, 22).

[17] E. Kahya and J. A. Dracup. “U.S. Streamflow Patterns in Relation to the El Niño/Southern Oscillation.” In: *Water Resources Research* 29.8 (1993), pp. 2491–2503 (cit. on pp. 1, 11).

[18] A. S. Kiem and S. W. Franks. “On the identification of ENSO-induced rainfall and runoff variability: a comparison of methods and indices.” In: *Hydrological Sciences Journal* 46.5 (2001), pp. 715–727 (cit. on pp. 11, 12).

[19] F. Li and Q. Zeng. “Statistical Prediction of East Summer Asian Monsoon Rainfall Based on SST and Sea Ice Concentration.” In: *Journal of Meteorological Society of Japan* 86.1 (2008), pp. 237–243.

[20] R. J. May et al. “Non-linear variable selection for artificial neural networks using partial mutual information.” In: *Environmental Modelling & Software* 23.10-11 (2008), pp. 1312–1326 (cit. on p. 26).

[21] E. Meidani and S. Araghinejad. “Long-Lead Streamflow Forecasting in Southwest of Iran by the Sea Surface Temperature of Mediterranean Sea.” In: *Journal of Hydrologic Engineering* (2013) (cit. on p. 1).

[22] I. D. Phillips and J. Thorpe. “Icelandic Precipitation - North Atlantic Sea-Surface Temperature Associations.” In: *International Journal of Climatology* 26 (2006), pp. 1201–1221 (cit. on pp. 2, 13).

[23] X. Quach. “Assessing and optimizing the operation of the HoaBinh reservoir in Vietnam by multi-objective optimal control techniques.” PhD Thesis. Politecnico di Milano, Milan, Italy, 2011 (cit. on pp. 33, 35).

[24] T. A. Räsänen and M. Kummu. "Spatiotemporal influences of ENSO on precipitation and flood pulse in the Mekong River Basin." In: *Journal of Hydrology* 476 (2013), pp. 154–168 (cit. on p. 3).

[25] O. Roswintiarti, S. N. Devdutta, and S. Raman. "Teleconnectionsb etweent ropical Pacific sea surface temperature anomalies and North Carolina precipitation anomalies during El Nifio events." In: *Geophysical Research Letters* 25.22 (1998), pp. 4201–4204 (cit. on pp. 2, 4, 22).

[26] N. Savage. "Modelling: Predictive yield." In: *Nature* 501.7468 (2013), pp. 10–11 (cit. on p. 1).

[27] A. Sharma. "Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - A strategy for system predictor identification." In: *Journal of Hydrology* 239.1-4 (2000), pp. 232–239 (cit. on p. 22).

[28] A. Shrestha and R. Kostaschuk. "El Niño/Southern Oscillation (ENSO)-related variablity in mean-monthly streamflow in Nepal." In: *Journal of Hydrology* 308.1-4 (2005), pp. 33–49 (cit. on p. 11).

[29] T. L. Soukup et al. "Long lead-time streamflow forecasting of the North Platte River incorporating oceanic–atmospheric climate variability." In: *Journal of Hydrology* 368 (2009), pp. 131–142 (cit. on p. 1).

[30] A. O. Tarakanov and A. V. Borisova. "Galapagos indicator of El Niño using monthly SST from NASA Giovanni system." In: *Environmental Modelling & Software* 50 (2013), pp. 12–15 (cit. on pp. 2, 12).

[31] L. Yungang, H. Daming, and Y. Changqing. "Spatial and temporal variation of runoff of Red River Basin in Yunnan." In: *Journal of Geographical Sciences* 18 (2008), pp. 308–318 (cit. on p. 48).