THESIS

STEPWISE NONPARAMETRIC DISAGGREGATION FOR DAILY STREAMFLOW GENERATION CONDITIONAL ON HYDROLOGIC AND LARGE-SCALE CLIMATIC SIGNALS

Submitted by

José Manuel Molina Tabares

Department of Civil and Environmental Engineering

In partial fulfillment of the requirements

for the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2010

COLORADO STATE UNIVERSITY

April 2nd, 2010

WE HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER OUR SUPERVISION BY JOSE MANUEL MOLINA TABARES ENTITLED STEPWISE NONPARAMETRIC DISAGGREGATION FOR DAILY STREAMFLOW GENERATION CONDITIONAL ON HYDROLOGIC AND LARGE-SCALE CLIMATIC SIGNALS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

Committee on Graduate Work

_____
Stephanie K. Kampf

_____
David A. Raff

_____
Advisor: Jorge A. Ramírez

_____
Co-Advisor: José D. Salas

_____
Department Head: Luis García

ABSTRACT OF THESIS

STEPWISE NONPARAMETRIC DISAGGREGATION FOR DAILY STREAMFLOW GENERATION CONDITIONAL ON HYDROLOGIC AND LARGE-SCALE CLIMATIC SIGNALS

A stepwise nonparametric stochastic disaggregation framework to produce synthetic scenarios of daily streamflow conditional on volumes of spring runoff and large-scale ocean-atmosphere oscillations is presented. This thesis examines statistical links (*i.e.*, teleconnections) between decadal/interannual climatic variations in the Pacific Ocean and hydrologic variability in US northwest region, and includes a spectral analysis of climate signals to detect coherences of their behavior in the frequency domain. We explore the use of such teleconnections of selected signals (e.g., north Pacific gyre oscillation, southern oscillation, and Pacific decadal oscillation indices) in the proposed data-driven framework by means of a cross-validation-based combinatorial approach with the aim of simulating improved streamflow sequences when compared with disaggregated series generated from flows alone. A nearest neighbor time series bootstrapping approach is integrated with principal component analysis to resample from the empirical multivariate distribution. A volume-dependent scaling transformation is implemented to guarantee the summability condition. The downscaling process includes

a two-level cascade scheme: seasonal-to-monthly disaggregation first followed by monthly-to-daily disaggregation. Although the stepwise procedure may lead to a lack of preservation of the historical correlation between flows of the last day of a month and flows of the first day of the following month, we present a new and simple algorithm, based on nonparametric resampling, that overcomes this limitation. The downscaling framework presented here is parsimonious in parameters and model assumptions, does not generate negative values, and preserves very well the statistical characteristics, temporal dependences, and distributional properties of historical flows. We also show that both including conditional information of climatic teleconnection signals and developing the downscaling in cascades decrease significantly the mean error between synthetic and observed flow traces. The downscaling framework is tested with data from the Payette River Basin in Idaho.

José Manuel Molina Tabares
Department of Civil and Environmental Engineering
Colorado State University
Fort Collins, CO 80523
Spring 2010

ACKNOWLEDGMENT

I would like first to thank my advisor and professor, Dr. Jorge Ramírez, for his support and guidance during my studies. His valuable insights and commitment to this research are greatly acknowledged. I would like to thank the members of my thesis committee, Dr. Jorge Ramírez, Dr. José Salas, Dr. David Raff, and Dr. Stephanie Kampf, for their valuable comments and suggestions. Special thanks to my professor, Dr. José Salas, his sincere and wise advice has been and will always be very much appreciated.

I thank my friends at the Hydrology lab for their sincere friendship and company during my academic process.

I would also like to thank the U.S. Bureau of Reclamation Science and Technology Program and Colorado State University for the funding support provided to my studies and research.

My family deserves a very special acknowledgment, for supporting me and being there with me during this challenging journey.

Finally, I thank God, beginning and end of everything.

This work is dedicated to my loving family. To my wife Dagna, who always believed in me. To my son Nicholas and my daughter Catalina, source of my inspiration

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Stepwise Nonparametric Disaggregation for Daily Streamflow Generation conditional on Hydrologic and Large-Scale Climatic Signals

1 Introduction

Generation of synthetic sequences of daily hydroclimatic variables like streamflow is often used for efficient short-term and long-term planning, management and assessment of complex water resources systems (Rajagopalan et al., 2010). Downscaling[1] methods are an important component of the hydrologist's tool kit for generating such flow traces, which should be statistically indistinguishable from the observations. Disaggregation of annual (or seasonal) runoff volumes to daily flows represents a difficult task on the one hand because it involves a large increase in the dimensionality of the problem (as compared to disaggregation from annual to monthly flows, for example) and on the other because daily flows exhibit nonlinear dynamics and present multimodal behavior with modes not normally distributed. Therefore, the functional structure associated to generation of daily streamflows does not lend itself easily to the well-known parametric approaches. For example, models to downscale to daily time scales based on parametric approaches do not properly account for the rising limb and recession characteristics that are typical of hourly and daily flow hydrographs (Rajagopalan et al., 2010). On the other hand, nonparametric methods require only a limited set of assumptions about the structure of the data, and they may therefore be preferable when *a priori* postulations required for parametric models are not valid (Efron

1 Downscaling has often been referred to as disaggregation in the past. In this document, we will use both terms indistinguishably

and Tibshirani, 1993; Higgins, 2004). In this thesis, we develop new, nonparametric-based tools for the downscaling of seasonal runoff volumes to daily flows. The new approach extends previous works on nonparametric modeling with the aim of developing a stochastic framework focused on temporal downscaling of seasonal volumes to daily values at a single site. The proposed methodology is implemented for the Payette River Basin in Idaho, where the streamflow dynamics is dominated by the dynamics of snow accumulation and melt.

Disaggregation provides synthetic realizations of the components of a vector of any variable **X,** given an aggregate variable **Y.** On a river section, for example, the streamflow disaggregation approach can be considered as simulating from the conditional probability density function (PDF), **f(X|Y)**, in which **X** is a vector that contains disaggregate values (e.g., daily volumes), and **Y** represents the aggregate amount (e.g., a seasonal volume), subject to the condition that the disaggregate values add up to the aggregate value $y_p = x_{1p} + x_{2p} + \ldots + x_{np}$. The subscript n is the size of **X** (e.g., n is the number of days in the season), and p relates to any specific period, season or year. In our case study, **Y** contains the aggregate value of flow at a single station and, as will be explained later, other climatic variables that are correlated with hydrologic dynamics in the study area. Although our case study deals only with a single site, in general, the vector **Y** may contain aggregate values of flow at a set of sites. In that case, the vector **X** would contain the disaggregate values at each site. The goal of the downscaling procedure is to generate time series of downscaled values at any specific site such that they exhibit patterns, statistical characteristics, temporal dependences, and distributional properties that are statistically indistinguishable from those of the observations. The

model's ability to generate such patterns is measured with cross-validation techniques. In order to check the preservation of statistical features, the process of stochastic modeling commonly uses synthetic generation by means of Monte Carlo simulation (Bras and Rodriguez-Iturbe, 1985).

Both parametric and nonparametric methods have been widely used for stochastic simulation and forecasting of hydroclimatic processes, the former over the past 40 years, the latter more recently during the last decade or so. Although both types of approaches have relative advantages and disadvantages, one approach or the other may be more appropriate for a specific situation. In the next paragraphs we highlight a few of the disadvantages of the parametric methods in the context of downscaling seasonal flows to daily flows, as well as the advantages that non-parametric methods offer in that context. However, for a general review of advantages and disadvantages of both approaches, the reader is kindly referred to Rajagopalan et al., (2010). Parametric stochastic models have long been used for streamflow disaggregation, generally from annual to seasonal timescales (e.g., annual to monthly) (e.g., Harms and Campbell, 1967; Valencia and Schaake, 1973; Mejia and Rousselle, 1976; Lane, 1979; Salas et al., 1980; Stedinger and Vogel, 1984; Stedinger et al., 1985; Grygier and Stedinger, 1988; Santos and Salas, 1992; Koutsoyiannis and Manetas, 1996). The evolution of the development of such schemes was driven by the need for improving model performance (e.g., reproduction of historical sample attributes) and increasing the model flexibility and parsimoniousness (e.g., by reducing parameterization), as well as to ensure conservation of mass (i.e., the *so-called* summability condition requiring that the seasonal volume implied by the daily disaggregated flows equals the seasonal volume observed) and avoiding generation of

negative values among others. Most of these drawbacks are commonly associated with distributional assumptions, transformations of the original data, and linear fittings not always valid (Tarboton et al., 1998). For example, in the traditional simulating process applying autoregressive (AR) and moving average (MA) approaches (e.g., ARMA(p,q), PARMA(p,q), etc.) to skewed data, a linear regression model based on a conditional expectation with and added random innovation is fit to the Gaussian transformation of the original series (e.g., Salas, 1993). These models are equivalent to simulating from a Gaussian conditional PDF because they assume the data are normally distributed and thus, the process of stochastic generation will produce also a normally distributed variable (Hipel and McLeod, 1994). Therefore, such disaggregation model preserves observational statistics only in the transformed (i.e., Gaussian) space, but the backtransformation of generated traces to the real flow domain does not guarantee preservation of observed statistics or conservation of mass (Tarboton et al., 1998). An alternative approach that can circumvent this problem is the so-called Periodic Gamma AR model (PGAR), which has been applied successfully to weekly streamflow generation for several rivers in the United States (Fernádez and Salas, 1986; Rajagopalan et al., 2010). However, the restrictions in using parametric approaches become more evident as the disaggregation timescale gets finer, as mentioned previously (e.g., annual to seasonal vs. seasonal to daily), or when the streamflow dynamics exhibit an intermittent behaviour (Rajagopalan et al., 2010), although periodic product models have been proposed for simulating monthly flows of some ephemeral streams (Chebaane et al., 1995).

Although nonparametric techniques have their own disadvantages (e.g., Rajagopalan et al., 2010), they have been proposed (e.g., Lall and Sharma, 1996), in part to overcome some of the shortcomings of parametric approaches in stochastic hydrology. Nonparametric approaches not only reduce, or even suppress parameterization and avoid assumptions about the distributional properties of the data, but also allow assessment of the variability of the quantities of interest without long-winded and error-prone analytical calculation (Davidson and Hinkley, 1997). Also, nonparametric techniques estimate the PDF directly from the observational data and have the ability to approximate specific features of the underlying true density (multimodality, non-zero skewness, etc.) (Simonoff, 1996; Takezawa, 2006). Recent works on streamflow disaggregation based on nonparametric methods have shown important improvements over the traditional parametric models (e.g., Tarboton et al., 1998; Kumar et al., 2000; Sharma and O'Neil, 2002; Prairie et al., 2007; Lee et al., 2010). However, some non-parametric approaches still have some shortcomings common to parametric schemes, such as generation of negative flows in some cases, lack of preservation of cross-boundary correlations (e.g., the model by Prairie et al. (2007) was unable to capture the month-to-month correlation between December of the previous year and January of the current year) and generation of flow sequences that can be very similar to the observations (Rajagopalan et al., 2010; Lee et al., 2010). Data-driven approaches, including some hybrid methods that combine parametric and nonparametric techniques (Srinivas and Srinivasan, 2005), have been focused on time and/or space disaggregation and, as in the parametric case, implemented generally from annual to monthly timescales. The exception is the work of Kumar et al. (2000), in which the disaggregation framework, based on a constrained optimization

function, was developed to disaggregate monthly streamflows to daily values. However, the Kumar et al. (2000) approach leads to a single deterministic solution because no stochastic scheme was included in their algorithm, and thus, their model's ability to simulate the random nature of the streamflow series is limited.

The disaggregation problem requires the definition of the distributional forms for the joint and marginal probability densities of the flow variables, from which the conditional probability density function is derived. Tarboton et al. (1998) and Sharma and O'Neil (2002) implemented kernel density estimation (KDE) for the estimation of the conditional and joint PDFs of the flow variables. Prairie et al. (2007) simulate from the conditional PDF, but replacing KDE with a K-nearest-neighbor (KNN) resampling scheme (Lall and Sharma, 1996), taking into account that kernel methods can be limited and may not work properly where many variables are being modeled simultaneously (Simonoff, 1996; Sharma and O'Neil, 2002). Because the accuracy of nonparametric methods depends on the sample size, KDE needs progressively larger sample sizes in higher dimensions to achieve comparable accuracy with respect to lower dimensions. This is the result of the so-called *empty space phenomenon* (Simonoff, 1996), which refers to the fact that in higher dimensions (i.e., many variables being modeled), large neighborhoods in the data space will likely have sparse observations leading to the loss of the local character.

Tarboton et al. (1998), Prairie et al. (2007) and Bracken et al. (2010) use Gram-Schmidt (G-S) orthonormalization in order to specify $\mathbf{f(X|Y)}$ and to ensure the mass conservation of disaggregate to aggregate flows. However, the G-S procedure not only may produce negative values in the disaggregated traces after back-rotation (mainly in

low-flow seasons), but also misrepresentation of long-term statistics such as the variance and minimum values (mainly in the case of ephemeral streams). Tarboton et al. (1998) and Prairie et al. (2007) report generation of negative monthly flows in the locations evaluated. Tarboton et al. (1998) attributed this to the KDE used in their disaggregation approach and they incorporated additional algorithms for overcoming such problem. Prairie et al. (2007) stated that this negative generation had no significant impacts on their model evaluation. Additionally, the G-S transformation may be computationally demanding in a higher dimensional disaggregation problem (e.g., annual to daily). A volume-dependent scaling transformation of time series is used in this work as a simple alternative to the G-S procedure in order to both avoid negative generation and guarantee conservation of mass. Therefore, the volume-dependent scaling transformation is also an alternative to implement streamflow disaggregation in ephemeral streams.

All currently available nonparametric disaggregation models have considered only observational flows in the definition of the conditional distributions. However, growing evidence indicates the importance of the role of the thermal inertia of the world's oceans in determining the patterns of global atmospheric water and energy flows, and consequently, hydrologic variability. Therefore, the identification and characterization of coupled ocean-atmosphere interactions plays a key role in explaining the variability of hydrologic processes like streamflow. In the western United States, for example, large-scale climate phenomena like El Niño-Southern Oscillation (ENSO), Pacific Decadal Oscillation (PDO) and others have shown strong links with the hydroclimatology of the study region (e.g., Mantua et al., 1997; Gershunov, 1998; McCabe and Dettinger, 1999; Hamlet and Lettenmaier, 1999; Clark et al., 2001;

Harshburger et al., 2002; Hidalgo and Dracup, 2003; Barton and Ramirez, 2004; Grantz et al., 2005; Regonda et al., 2006; Sobolowski and Frei, 2007; Linkin and Nigam, 2008; Bracken et al., 2010), and in some cases, specific climatic interactions have been used for improving the skill of seasonal streamflow forecasts (e.g., Hamlet and Lettenmaier, 1999; Hamlet et al., 2002; Grantz et al., 2005; Regonda et al., 2006; Bracken et al., 2010). These links (also referred to as teleconnections) between large-scale atmospheric or oceanic phenomena and remote local or regional hydroclimatic responses are commonly defined by lagged statistical associations along with specific persistence characteristics. Bearing this in mind, the work reported here also considers time and spatial cross-covariance analyses between seasonal streamflow and large-scale climatic variables to search and identify specific climatic patterns, periods and potential regions in the Pacific Ocean that exhibit statistically significant teleconnections with the hydrological response in the study region.

This thesis presents a new stepwise nonparametric multivariate conditional disaggregation model. For future reference, the proposed model is termed MuNDi-S for Multivariate Nonparametric Disaggregation. Relevant concepts developed by Lall and Sharma (1996) in the stochastic hydrology field are utilized in this model. Regarding the application of nonparametric methods to disaggregation of seasonal streamflow, this work makes important new contributions that extend those of Tarboton et al. (1998) and Prairie et al. (2007). In addition to incorporating information on ocean-atmosphere oscillations in the input vector, our stochastic nonparametric framework is developed to disaggregate to a smaller time scale, from seasonal to daily values, using a cascade (piece-wise) scheme. The MuNDi-S model disaggregates to daily time scale conditional

on forecasted spring runoff volumes and observed states of the ENSO, PDO, and the North Pacific Gyre Oscillation (NPGO). Recent evidence suggests that the NPGO is physically connected to tropical variability through the ENSO dynamics (Di Lorenzo et al, 2009). An additional important aspect of the new data-driven model is that it combines the mentioned strengths of nonparametric techniques with those of multivariate methods for extracting important information from complex data sets. A KNN time series bootstrap scheme (Lall and Sharma, 1996) is integrated with Empirical Orthogonal Function (EOF) analysis and the related Principal Components Analysis (PCA) for the development of our disaggregation framework. PCA involves a mathematical procedure that transforms a set of correlated response variables into a smaller set of uncorrelated variables called *principal components* (PC) (Johnson, 1998). One important motivation for incorporating selected teleconnection signals and/or implementing a stepwise scheme in MuNDi-S comes from a desire for generating improved streamflow traces when compared with disaggregated series generated from flows alone and/or without cascades. A cross-validation-based combinatorial algorithm was developed for testing purposes. KDE is also considered for the definition of probability distributions in the model evaluation. The MuNDi-S approach can be extended in future works to space-time disaggregation.

This document is organized as follows. Section 2 presents a brief description of the study region and introduces the hydroclimatic information used. Section 3 describes the new downscaling methodology. Section 3 first introduces the procedure for identification of teleconnections between large-scale climatic phenomena and the hydrology of the study region as well as the procedure for determining periodicities of

climate signals. This is followed by the proposed model and algorithms for daily streamflow generation conditional on streamflow and climate fluctuations. In section 4, the disaggregation framework is evaluated by applying it to the flows of the Payette River Basin, Idaho. Section 4 also presents statistical and spectral analyses of the climate signals considered. Section 5 presents a summary and final remarks.

2 Study Region and Data

2.1 River Basin and Hydrology

Daily streamflow records for the period 1950-2007 were obtained from the USBR Pacific Northwest region database (http://www.usbr.gov/pn/hydromet/). The study region in the Payette River Basin is shown in Figure 1. The historical streamflow records correspond to the USBR gauging station *Payette River near Horseshoe Bend*, code HRSI, latitude 43°56'42", longitude 116°11'49" and elevation 2626 feet. Regulated streamflow locations are avoided in this study because the observed streamflow is significantly altered by human activity (e.g., irrigation diversions, reservoir releases, etc.). The dataset in our analyses corresponds to a river section with unregulated flow. Spring seasonal and monthly volumes (April – July) for the period 1950 – 2007 are computed from daily records. Precipitation in the basin generally accumulates as snowpack during the winter (December through March), and provides a snowmelt pulse in the spring. The inset plot in Figure 1 shows the long-term periodic mean and standard deviation of monthly streamflows for the period of analysis. A strong seasonality in the mean flow is observed, with low base flows from August through February, a rise in early spring (March-April),

peaks in May and June and a recession in late summer. The most important runoff volume occurs in the period April-July. Patterns of the periodic mean and the periodic standard deviation curves are in phase, meaning that there is a direct relation between the dispersion of flow data and its magnitude.

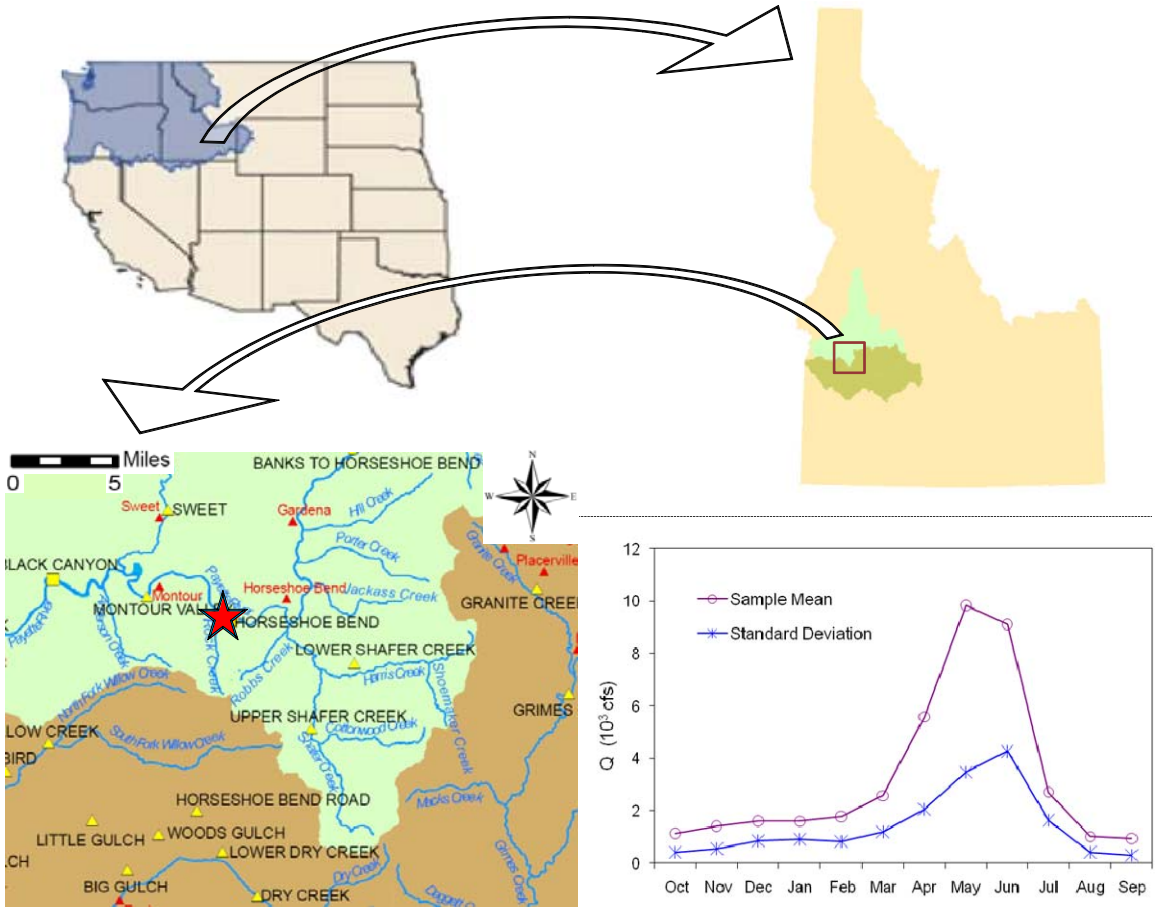

Figure 1. Location of the Payette River Basin (green area) in Idaho, and the USBR gauging station *Payette River near Horseshoe Bend* (red star), code HRSI, latitude 43°56'42", longitude 116°11'49". The long-term periodic mean and standard deviation of monthly streamflows are shown in the inset.

2.2 Large-Scale Climatic Information

Analyses of observational sea surface temperature (SST) and sea level pressure (SLP) data are of primary importance for understanding both global climate and climatic phenomena and their relationships with local hydroclimatological processes. Historical records on SST and SLP fields were used in this work to explore statistical links between

both high- and low-frequency climatic variations in the Pacific Basin and the streamflow variability in the study region. Monthly SST and SLP fields on a 5°x5° latitude/longitude resolution in the Pacific Ocean are available from the National Oceanic and Atmospheric Administration (NOAA) website, Physical Sciences Division (PSD) of the Earth System Research Laboratory (NOAA/ESRL) (http://www.cdc.noaa.gov).

Climatic feedbacks involving SST and SLP anomalies play an important role in defining the ENSO phenomenon (Trenberth, 1997), and they have been shown to be one of the earth's most important ocean-atmosphere interactions. The dynamics and impacts of ENSO around the world have been well documented in the last three decades, although its causes are complex and remain not well understood (Wang C. et al., 2004). The term El Niño (EN) refers to extended and periodic episodes of anomalous warming of the Pacific Ocean (specifically, in SSTs) off the coast of Peru (between approximately the date line and 120° W). Southern Oscillation (SO) is the atmospheric component of ENSO and refers to a SLP swing between the locations Darwin in Australia, and the island of Tahiti in the central tropical Pacific (Figure 2). SSTs and SLPs are strongly correlated in the study of ENSO, although EN and SO are two different but related aspects of this phenomenon. The gradient defined by the SO in the form of a standardized index (i.e., the Southern Oscillation Index, SOI) has been commonly used to measure the magnitude of an ENSO event (Trenberth, 1997). A warm phase of ENSO (i.e., EN) occurs about every 3-7 years and is characterized by high negative values of SOI, and alternated with high positive SOIs (La Niña). Although several large-scale climate features associated with ENSO play an important role in modulating the interannual variability of western U.S. hydroclimatology, the SOI was used in this research as a measure of ENSO activity.

SST-based ENSO patterns like Niño 1+2, Niño 3, Niño 4, Niño 3.4, etc., can also be easily incorporated and assessed in the framework developed here. The seasonal average values of the SOI in the period 1950-2007 are shown in Figure 2. Monthly data on the SOI are available from the website of the US National Weather Service (NOAA/NWS), at http://www.cpc.ncep.noaa.gov/data/indices/.

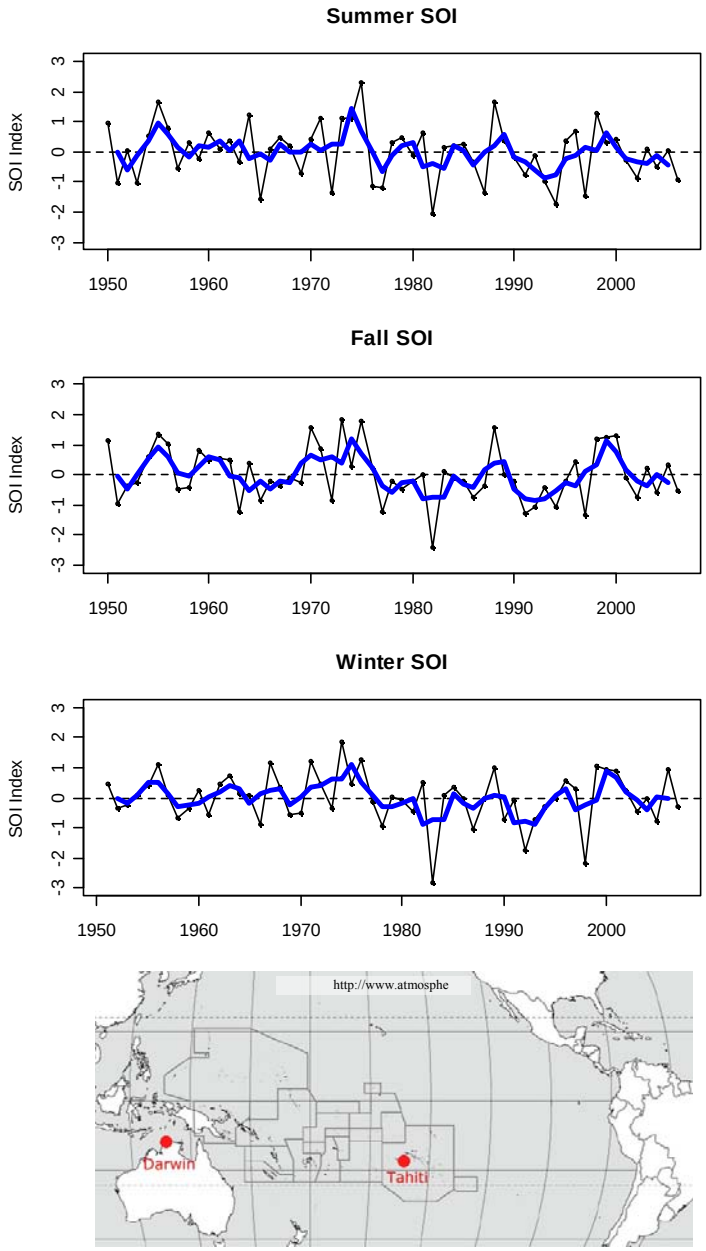

Figure 2. SOI time series for the summer (July-September), fall (October-December), and winter (January-March) seasons. The period of analysis is 1950–2006 for summer SOI and fall SOI, and 1951–2007 for winter SOI. The blue line corresponds to the smoothed series using a three-year moving average function. The lower plot shows the locations of Darwin and Tahiti in the tropical Pacific Ocean used to calculate the SOI.

The PDO is the main oceanic expression of climatic variability in the North Pacific Ocean. The PDO refers to a warming or cooling of the surface waters in this region, north of 20° N, and fluctuations in SSTs for this phenomenon are detected mainly in a decadal time scale. Further details of this climatic variation are presented by Mantua et al. (1997), Wang D. et al. (2004) and Chhak et al. (2009). The PDO presents positive phases determined by below normal SST anomalies in the central North Pacific Ocean. Recent findings support the hypothesis that the tropics play a key role in the North Pacific climate variability (Deser et al., 2004; Schneider and Cornuelle, 2005; Di Lorenzo et al, 2009), and there exists evidence that suggests the notion that the PDO phasing modulates the hydroclimatological impacts associated with ENSO events in the western U.S. (see Gershunov and Barnett, 1998; Koch and Fisher, 2000; Harshburger et al., 2002; Barton and Ramirez, 2004). However, there is still controversy about whether the PDO is truly independent of the ENSO pattern (Newman et al., 2003). Schneider and Cornuelle (2005) suggested that in North America the stratification of climate anomalies or teleconnection patterns should be based on the underlying indices of ENSO and the North Pacific Index (NPI) (as defined by Trenberth and Hurrell, 1994), rather than ENSO and the PDO. The PDO index is defined as the first principal component (PC1) of the SST field over the given region. The PDO index computation includes monthly series of SSTs on a 5°x5° latitude/longitude grid, north of 20° N in the Pacific Ocean. Long-term monthly anomalies are then calculated (i.e., mean removal). Following, the global warming SST trend is removed from the long-term anomalies at each grid point/month (i.e., variance removal). EOFs are then computed for each month in the grid and the first principal component is retained (Mantua et al., 1997). Long-term time series of the PDO

index are available from the website of the Joint Institute for the Study of the Atmosphere and Ocean (JISAO), University of Washington, at http://jisao.washington.edu/data_sets/. The July-September (summer), October-December (fall), and January-March (winter) average values of the PDO index series in the period 1950-2007 are plotted in Figure 3. An interdecadal oscillation is apparent in the smoothed time series.

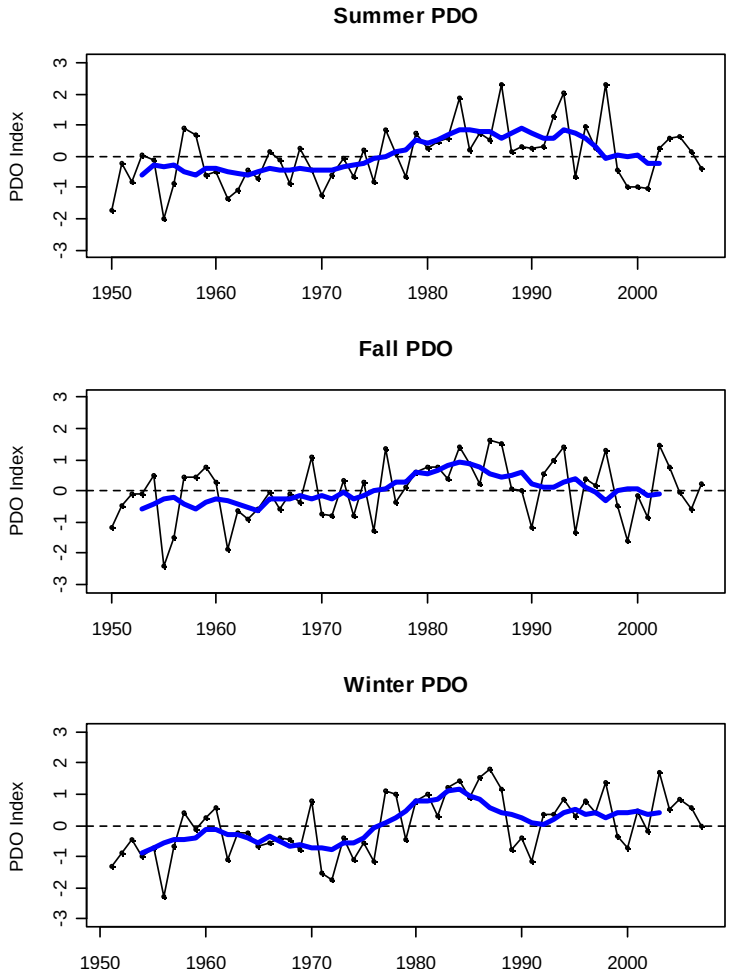

Figure 3. Same as Figure 2 but for PDO index time series. The smoothed lines using an eight-year moving average function. An interdecadal oscillation of this phenomenon is apparent, and a change in polarity in 1977 from negative to positive is observed.

However, the mechanisms that couple fluctuations in the PDO and ENSO to changes in physical variables in the North Pacific Ocean remain unclear, and in many cases such fluctuations in the PDO and ENSO fail to explain decadal variations of key climatic and biological variables in the Northeast Pacific Ocean (Schneider et al., 2005; Di Lorenzo et al., 2005, 2008). Recent studies have identified a new pattern of climate variability, the North Pacific Gyre Oscillation (NPGO), defined as the second leading empirical orthogonal function of the Sea Surface Height (SSH) anomalies in the North Pacific Ocean (Di Lorenzo et al., 2008). These researchers show that the NPGO variability is significantly correlated with previously unexplained fluctuations of environmental variables in the Northeast Pacific Ocean, and show that this climatic pattern is able to explain interannual and decadal variations of salinity, nutrient upwelling and surface chlorophyll-a (Chl-a) in such oceanic region. The fluctuations in the NPGO index reflect changes in the intensity of the North Pacific gyre circulations (Di Lorenzo et al, 2008). The PDO and NPGO are the oceanic expression of the two dominant modes of North Pacific atmospheric variability, the Aleutian Low (AL) and the North Pacific Oscillation (NPO), respectively (for further details on how atmospheric forcing drives the low-frequency oceanic variability of the PDO and NPGO, readers are referred to Chhak et al., 2009). The AL and NPO are the first and second modes of SLP variability in the North Pacific basin, respectively, and have been associated along with ENSO to variations in historical patterns of climate over North America (Linkin and Nigam, 2008; and references therein). The former research suggests that the NPO fluctuations are likewise more influential than those of ENSO or the Pacific North American Oscillation (PNA) (a decadal atmospheric pattern commonly associated with the PDO signatures) on

the winter precipitation in the continental Pacific Northwest U.S. Recently, Di Lorenzo et al. (2009) proposed a conceptual framework that links such oceanic low-frequency variations in the North Pacific (i.e., PDO, NPGO), and their atmospheric forcings (e.g. AL, NPO), with the ENSO fluctuations. In a similar way to the computation of the PDO index, principal component analysis is used to define the NPGO index. The latter represents the second PC of the SSH anomalies over the region 180°W - 110°W and 25°N - 62°N, and very closely tracks the second dominant mode of the North Pacific SST field (the Victoria mode). Because SSTs and SSHs are highly correlated in the mentioned region ($R \approx 0.85$ between the first PCs of both fields for the period 1950 - 2007), Di Lorenzo et al. (2008, 2009) and Chhak et al. (2009) redefined the PDO index in terms of the SSH anomalies, computing its first PC instead of the first PC of the SST anomalies. Thus, the PDO index defined in this way and the NPGO index are conceptually uncorrelated series. The summer, fall, and winter mean NPGO index values in the period 1950-2007 are plotted in Figure 4. The smoothed NPGO observed in Figure 4 exhibits higher low-frequency oscillations than those of the PDO (this finding is further discussed with spectral analysis, section 4), which may be very helpful in finding relationships with decadal/interannual variability of many hydroclimatic and biological variables in the land-ocean system. Monthly averages of the NPGO index are available from the website http://www.o3d.org/npgo/data/NPGO.txt/. Given the significant statistical relationships found by Di Lorenzo et al. (2008) between environmental variables in the Northeast Pacific Ocean and the NPGO, in addition to the PDO and ENSO, we consider also the NPGO regarding hydroclimatic variability over the western U.S. Although information of ENSO in terms of the SOI is available from 1882, and of PDO is available from 1900, the

historical period of climatic indices considered in the analysis presented here is 1950 – 2007, because both the streamflow series in the study region and the NPGO records are available from 1950 only.

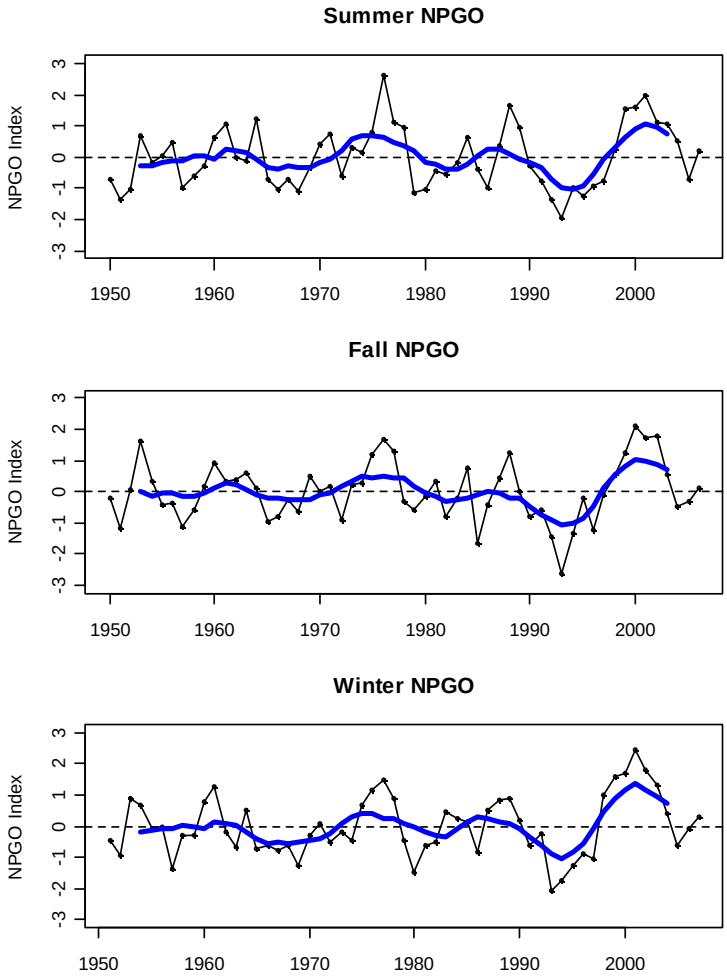

Figure 4. Same as Figure 2 but for NPGO index time series. The smoothed lines using an eight-year moving average function. Apparently, the NPGO low-frequency fluctuation about the neutral condition (index = 0) observed in the smoothed series is higher than the one of the PDO shown in Figure 3. Between four and five cycles in the period of study can be observed here, with the last change in polarity around 1997. An increased intensity of the NPGO over the last decade can be observed.

3 Methods

3.1 Identification and Statistical Analysis of Teleconnections

Spatial cross-covariance analyses between large-scale climatic patterns and seasonal streamflow were performed in order to define and quantify potential teleconnections. Such climatic patterns can be defined in terms of the variable type (e.g., SST, SLP, SSH, geopotential height, etc.), the preceding period (e.g., a 3-consecutive-month period) to the runoff season and the type of aggregation. We searched and identified climatic fluctuations based on SST and SLP fields in the Pacific Ocean that potentially account for the variability of spring (April through July) runoff volumes in Payette. In this statistical analysis, specifically, the period of climatic variables leading the spring runoff was established as three consecutive months, and the type of aggregation as the mean of the three-month period.

Many authors have determined that the effects of Pacific climatic phenomena are highly persistent and have a long-range effect in the hydrology of the western U.S. (see Hamlet and Lettenmaier, 1999; Harshburger et al., 2002; Grantz et al., 2005; Regonda et al., 2006; Bracken et al., 2010). Monthly SST and SLP fields averaged over three-consecutive-month periods were considered in this analysis, with the moving averages leading the spring runoff during the preceding summer, fall and winter seasons. The preceding three-consecutive-month averaged SST and SLP series were statistically associated with historical series of spring runoff volumes in the Payette River Basin. The Pearson correlation coefficient (R) was used as metric in the spatial analysis. A p-value of 0.05 (2-sided) of the T-distribution test for the correlation was used to define its statistical

significance (significance level = 95%). Based on results presented in section 4, the SST zones in the Pacific Ocean with the strongest correlation with spring runoff volumes in the study area correspond to geographical regions where PDO and ENSO are defined. Therefore, PDO and SOI are selected as variables to include in our teleconnection analysis and in the disaggregation framework. Given that the SST and SSH fields are highly correlated in the North Pacific Ocean (Di Lorenzo et al., 2008; Chhak et al., 2009), and considering the background reviewed in the previous section, the NPGO index is also included both in the teleconnections analyses and in the proposed disaggregation framework.

Lagged teleconnections between spring runoff volumes and SOI, PDO, and NPGO indices from the preceding summer, fall and winter seasons are also analyzed by means of cross-correlogram functions. Series of three-consecutive-month averages of SOI, PDO, and NPGO indices in the period 1950-2007 are evaluated in the correlograms. A significance level of 95% is used in this analysis to define the significance of the correlation.

Pairwise scatterplots and correlation matrices are also constructed to analyze the relationships between spring runoff and those three-month averages of climatic indices that exhibit the best correlations with the runoff. The smoothed curves depicted in the scatterplots (Figure 10, lower panel) were fitted nonparametrically using Loess (Cleveland et al., 1992; Loader, 1999), a robust locally weighted polynomial regression. The fact that the fitting is done locally (e.g., in a bivariate space (u_j, w_j), $j=1,\ldots, N;$ with u_j as predictor) means that the fit at u_j (i.e., the expected value of w_j conditional on u_j, $E[w_j \mid u_j]$) is made using points in a neighborhood of u_j and it provides the capability to

capture nonlinearities present in the data. The span of each neighborhood (smoothing factor) is chosen so that the neighborhood contains a specified percentage of the data points. Given a point p in the conditional space, the fitted value at u_p is the value of a polynomial fit (linear or quadratic) to the data using weighted least squares, where the weight W_j for (u_j, w_j) in the neighborhood of u_p is estimated in such a way that W_j decreases as the distance from u_j increases. The Loess algorithm was performed with preselected arguments by: (1) using a first-order local polynomial, (2) smoothing with three fourths of the data, (3) reweighting with four iteractions to improve the robustness of the fit in the presence of outliers, and (4) fitting by means of a tricube kernel weight function (Simonoff, 1996).

3.2 Spectrum of Climate Signals

Large-scale climatic variations have an intrinsic cyclic behavior, and thus, they can be characterized by their dominant oscillation rate. Regarding the three-month averaged series of the SOI, PDO and NPGO patterns, moving average functions were used as a first attempt to approximate the main periodicity of the underlying phenomenon (see Figures 2 - 4). In this section a procedure to detect the various modes of fluctuation in the frequency domain of such climate signals is described. The aim of this analysis is to detect the regular behavior of the SOI, PDO and NPGO indices over time, by means of decomposing their empirical series into their regular components: periodic fluctuations and random noise. Such fluctuations can be expressed in terms of Fourier expansions. This spectral analysis is used here for determining the predominant frequencies of

oscillation by means of the discrete Fourier transform (DFT). The DFT is a complex-valued weighted average of the data given by

$$d(j/n) = n^{-1/2}\sum_{t=1}^{n} x_t \exp\left(-2\pi i^* t j/n\right) \qquad (1)$$

where n is the length of data or sample size, j is the number of cycles ($j=0, 1,..., n/2$), x_t is the discrete function defined by the climatic empirical series ($x_1,..., x_n$), i^* is the imaginary unit, and values j/n are the fundamental frequencies (also known as the Fourier frequencies). The scaled periodogram can be expressed as

$$P(j/n) = \left(\frac{2}{n}\sum_{t=1}^{n} x_t \cos\left(2\pi t j/n\right)\right)^2 + \left(\frac{2}{n}\sum_{t=1}^{n} x_t \sin\left(2\pi t j/n\right)\right)^2 \qquad (2)$$

Because the DFT is derived as a nonlinear regression of the climatic signal x_t on the sinusoids varying with values of j/n, regression coefficients need to be estimated at each Fourier frequency. To avoid computing repeated regressions, (2) was computed here more efficiently using a fast Fourier transform (FFT) algorithm (e.g., Shumway and Stoffer, 2006).

Any value of the periodogram given by (2) may be regarded as a measure of the squared correlation of the data with sinusoids oscillating at specific frequency j/n. The sum of all values $P(j/n)$ defined by (2) over the extent of fundamental frequencies accounts for the entire variability present in the underlying climatic oscillation. Because the periodogram quantifies the variance associated with each frequency, $P(j/n)$ can be thought of as the sample version of the spectral density, and here we refer to the $P(j/n)$ values as the sample spectral density or simply the spectrum. The frequency j/n is given

in cycles per time step of the observed records. Monthly series of the PDO (1900-2007), NPGO (1950-2007) and SOI (1950-2007) indices were used to define the periodograms.

3.3 General Description of the Disaggregation Framework

Disaggregation (more properly referred to as downscaling) is a procedure to obtain realizations of a given variable at a specific (time or space) scale conditional on realizations of the same variable at a larger scale, and such that the downscaled realizations are statistically indistinguishable from the population variable and meet the physical constraints of the specific variables (e.g., conservation of mass – sum of daily downscaled volumes over a season must equal the seasonal value). There exist many downscaling procedures that have been proposed in the literature, their differences depending on the specific assumptions that are made by each about the underlying nature of the variable being downscaled and on the specific application for which the procedure is being developed. As a result, many or all of the existing procedures have limitations in how well they meet the conditions of the general definition of downscaling presented above. In addition to limitations stemming from the assumptions made about the physical nature of the underlying processes (e.g., intermittent vs. continuous) those limitations are generally related to how well the downscaling procedure preserves the historical sample statistics of the underlying process. In addition, there are limitations having to do with the data requirements for parameter estimation and with mathematical tractability. In this thesis we propose a downscaling methodology based primarily on non-parametric approaches, as will be described next.

Our procedure for streamflow downscaling consists of scaling by sampling from the empirical conditional probability density function, **f(X|Y)**, i.e., the distribution of the observed **X** conditional on the observed **Y**. In our case, **X** is a vector of daily streamflows, and **Y** is a vector that contains the aggregate flow (seasonal volumes of runoff) and preceding averages of climatic indices, subject to the condition that the daily flows add up to the aggregate flow. Here, the conditional PDF is evaluated empirically using nonparametric techniques, as explained bellow. The conditional PDF **f(X|Y)** can be defined as

$$f(X/Y) = \frac{f(X,Y)}{f(Y)} = \frac{f(X,Y)}{\int f(X,Y)dX} \qquad (3)$$

where **f(X,Y)** is the joint PDF of **X** and **Y**, and **f(Y)** is the marginal PDF of **Y**. **f(Y)** can be obtained by integrating the joint PDF **f(X,Y)** over all the components of **X**. The joint and marginal functions are estimated from historical observations of **X** and **Y**. As indicated earlier, existing disaggregation models based on (3) have considered only streamflow information in the conditioning vector.

An important source of variability in the streamflow dynamics is the variability from regional climate, which is linked to large-scale oceanic and atmospheric variability. We hypothesize that incorporating specific large-scale climatic variables (or other hydroclimatic variables) in addition to seasonal flows in **Y** will lead to better preservation of statistical characteristics of **X,** and to more realistic and compatible streamflow patterns throughout the downscaled season. In this manuscript, seasonal volumes of runoff and combinations of seasonal averages of PDO, SOI and NPGO indices comprise the variables of the vector **Y**. In our case study, **X** is composed of daily volumes of

streamflow during the period of April to July. As stated earlier, an important condition that must be imposed on **X** is that of conservation of mass.

The general strategy for the development of the nonparametric disaggregation framework consists in a nearest neighbor resampling with adjustment of the historical time series, i.e., the resampling is done first followed by the adjustment based on a multiplicative scaling transformation as explained bellow. We use a KNN time series bootstrapping approach first, where a time series of daily flows on a period of interest (e.g., the four-month period April-July, or a month in that period) is resampled from the historical record by conditioning on a vector **Y** (hereafter, **Y** is also referred to as the *query point*) composed of the aggregate flow on that period of interest and averages of climate variables on preceding periods (e.g., the four-month period September-December of the previous year, or a month in that period). The time interval of the period of a climatic variable (i.e., 4 months or 1 month) matches that of the aggregate volume, as explained in section 3.4. Therefore, for every aggregate flow to be downscaled, the algorithm involves the identification of K vectors from the observational record (each k^{th} vector with the same variables and periods as in the query point) with the most similar hydroclimatic behavior to that of the query point. The identified K historical vectors are referred to as the "K nearest neighbors" of the query point. The degree of similarity between the nearest neighbors and the query point will be measured as a function of a modified Euclidian distance between the given vectors (see Section 3.4.1). In general, we use the hydroclimatic information given in the query point for identifying the K nearest neighbors within the N historical vectors that compose empirically **f(X,Y)**, where N is the sample size or number of years. Hereafter, these N historical vectors are referred to as the

"historical **Y**". The historical flow sequences from *K* identified nearest neighbors are then considered as potential vectors (in a stochastic sense and without previous transformations) to generate raw synthetic traces of daily flows. Therefore, the identification of these K-nearest neighbors plays a major role in the disaggregation scheme. Following in the simulation process, one of the KNN flow sequences is probabilistically selected for resampling and properly transformed while ensuring mass conservation. Such transformation based on a volume-dependent scaling also adapts to the serial dependence exhibited by historical data and guarantees that only positive flow values are generated in the disaggregated trace.

The disaggregation procedure consists of a two-level cascade in which a given seasonal volume (e.g., April-July runoff volume) is disaggregated into monthly values first (denoted as *first level*), and then each monthly volume is downscaled into daily components (denoted as *second level*). Because the downscaling is performed in two steps (or levels), each level requires the identification of its own set of nearest neighbors. In general, the identification of nearest neighbors depends on the disaggregation level. Thus, for the second level, the identification of *K* nearest neighbors is specific for each month, and at each month, one of its *K* nearest neighbors is probabilistically selected for resampling.

3.4 A Stepwise Multivariate Nonparametric Disaggregation Model - (MuNDi-S)

Considering the two disaggregation levels discussed previously, there are two types of periods of the variables that compose **Y**. A four-consecutive-month period is used for the first disaggregation level, and a one-month period is used for the second

disaggregation level. For the first disaggregation level, i.e., seasonal to monthly scales (or seasonal to daily scales if no stepwise is performed), **Y** is a vector whose number of components varies depending on the complexity of the model. In the simplest case, **Y** has only one component, Y_1, corresponding to the seasonal runoff volume; and in the most complex case, it has four components as follows, Y_1 is the spring (April-July) volume, and Y_2, Y_3, and Y_4 are the PDO, SOI and NPGO indices, each averaged over a 4-consecutive-month interval leading the spring runoff season (e.g., November-February, December-March, etc.). The vector **Y** always includes the seasonal runoff volume, but it also may contain one or more of the climatic variables, SOI, PDO, and NPGO. Several preceding periods of the averaged climatic indices are considered for evaluation of the MuNDi-S model and analyzed in section 4. The same vector array applies for the second disaggregation level, i.e., monthly to daily scales, but as previously mentioned, the components of **Y** corresponding to monthly values instead of seasonal values (e.g., Y_1 in April, and Y_2, Y_3, and Y_4 in preceding months). Note that for the first disaggregation level we select the time interval of the climatic index (i.e., 4 months) matching that of the aggregate volume because after the first disaggregation level, their corresponding monthly indices are used consecutively for downscaling consecutively monthly flows in the second disaggregation level. For example, if November-February PDO is used for disaggregating the spring volume into monthly values in the first disaggregation level, then in the second disaggregation level from monthly to daily scales, November PDO is used for daily disaggregation of the April volume, December PDO for daily disaggregation of the May volume, etc., i.e., a lag-time of five months between each monthly flow in the spring season and the corresponding monthly PDO index in the

November-February period is used in this specific example. In general, such approach allows consideration of the same lag-time between the underlying climate signal and the hydrologic response at each disaggregation level. Each of the steps in the disaggregation algorithm are discussed in more detail below, beginning with identification of the nearest-neighbors, followed by the multivariate decomposition analysis, series resampling, scaling transformation, concatenation of disaggregated daily traces, and serial dependence correction in the stepwise scheme.

3.4.1 Step 1: Identification of Nearest-Neighbors

The nearest neighbors are determined by estimating modified Euclidean distances in a *m*-dimensional space between the historical **Y** and the query point. This last one is a vector of scalars that contains as a first component the standardized value of the aggregate flow; the other *m-1* components are standardized values of climatic signals. All the components of the historical **Y** must be standardized also. Such transformation understood as centering the original data about the mean and reducing the centered data over the standard deviation. Standardization ensures that all elements of **Y** (historical and query) have commensurate variability, which is essential in the determination of the nearest neighbors. The Euclidian distance is evaluated as

$$\Delta_j = \left[\sum_{i=1}^{m} e_{i1} \left(query_i - obs_{ji} \right)^2 \right]^{\frac{1}{2}} \qquad for\ j = 1, 2, \ldots, N \qquad (4)$$

where $query_i$ is the i^{th} standardized component of the query point, obs_{ji} is the i^{th} standardized component of the j^{th} vector of the historical **Y**, and N is the number of years considered for analysis. The factor e_{i1} is used to assign a weight to each i^{th} squared

difference in (4). Such weighting is developed using principal component analysis (PCA) as explained next. Then, each one of the computed Δ_j corresponds to a PCA-weighted Euclidean distance averaged in a *m*-dimensional space. Scale-invariant Mahalanobis distances may be also used to determine the nearest neighbors. Such approach takes into account the correlation structure of the multivariate dataset, since it is calculated using the inverse of the covariance matrix. However, when the data contain much correlated information, the so-called multicollinearity problem, it may lead to a singular or nearly singular covariance matrix that cannot be inverted and, in that case, the Mahalanobis distance computation is truncated. The PCA-weighted Euclidean distance approach does not suffer from this limitation, and it also provides the capability to determine the nearest neighbors in those selected directions (PCs) where the highest variability is observed.

3.4.2 Step 2: Multivariate Decomposition Scheme with PCA

PCA transforms the historical **Y** into a set of independent variables or principal components, and each PC is expressed as a linear combination of weights (the so-called loadings) with variables from **Y**. Such loadings correspond to the components of the eigenvector associated with each PC, and each component in the eigenvector is associated with a variable of **Y**. Because each loading accounts for an amount of the variance explained by the PC, then, the loading is used here as a scaling weight of each variable of **Y** in the calculation of the Euclidean distance. The factors e_{i1} in equation (4) are estimated from the standardized data of the historical **Y**, and are equal to the absolute values of the loadings associated to the eigenvector of the first principal component (PC1) (this is the reason for using the subscript *1* in e_{i1}). By definition, the greatest

variance of **Y** lies on the first principal component. If additional PCs are retained, the absolute values of the loadings in each PC should be weighted by the normalized eigenvalue (from 0 to 1) of the respective PC, and factors $e_{i(\,)}$ are then calculated as the sum of weighted loadings over the specific components retained. Note that the eigenvalue is a measure of the variance of the data explained by its respective PC. If **Y** is a one-dimensional vector, PCA is not required and e_{i1} is equal to one.

The implementation of PCA in our disaggregation model has basically two major objectives in addition to avoiding collinearity issues: (1) dimensionality reduction in the number of variables of **Y** by retaining those transformed variables (lower-order PCs) that most contribute to the original variance of **Y**, and (2) extract important information (loadings) from the selected components in the rotated space. It is a normal practice in standardized PCA to retain the PCs associated with a variance value greater than 1.0. Such criterion is also known as the Kaiser-Guttman stopping rule, which is based on the average value of the eigenvalues (further details on stopping rules in PCA can be found in Jackson (1993), and many other sources). We select PC1 in the computation of Δ_j because in our case PCA results show that the Kaiser-Guttman rule is satisfied by this component only. If additional PCs contribute to the variance of **Y** almost as PC1 and satisfy the Kaiser-Guttman rule, they should be also retained. In that case, the factors $e_{i(\,)}$ must be computed as explained earlier. Our PCA is based on the covariance matrix of the standardized observational **Y**.

3.4.3 Step 3: Classification of KNN

The KNN bootstrapping scheme used in this research requires a previous definition of a neighborhood in the data space. Lall and Sharma (1996) suggest a prescriptive choice for the number of neighbors *K* as the square root of the sample size *N*. The Akaike information criteria (AIC) and generalized cross-validation (GCV) score functions can be also implemented in the estimation of *K* (Lall and Sharma, 1996; Rajagopalan and Lall, 1999). Following these works, we selected the size of the neighborhood as

$$K = \left(N - 1\right)^{0.5} \qquad (5)$$

and rounded to the nearest integer. Taking into account the magnitude of the PCA-weighted Euclidean distances defined by (4), KNN are identified from the *N* years evaluated, and organized by sorting values of Δ_j from the smallest to the largest one. A rank from 1 to *K* is assigned to the first *K* values of Δ_j. We refer to this classification from here on as PCA-KNN.

3.4.4 Step 4: Time Series Resampling

Disaggregation procedures aim for reproducing the historical autocorrelation in the generated traces. In order to preserve the serial dependence structure of the empirical data, a bootstrap technique is used over the identified PCA-KNN, where one of the nearest neighbors is selected for resampling. In general, resampling implies repeating an observed or estimated attribute of a variable to produce a synthetic realization of that variable. In particular, our resampling scheme selects the empirical monthly (or daily)

flow sequence x_{ip} (x_{ip}=x_{1p}, x_{2p}, ..., x_{np}; n is the size of the monthly (or daily) vector) of the selected neighbor in the period p (p =1, ..., P) (e.g., if an annual volume is disaggregated into monthly values, p =1, ..., 12) to be used as the sequence (before scaling) for the disaggregation period. The neighbor considered for resampling is randomly chosen according to a discrete resampling function $W(k, \Delta)$ that gives more weight to the closer nearest neighbors and less to the farther ones. The random selection of one neighbor from the PCA-KNN allows consideration of the random nature of the mechanism that generated the observations. In this way, we incorporate the uncertainty inherent in the realization of a flow trace, and weight the similarities that exist between the conditional information (query point) and the nearest neighbors. Here, we define $W(k, \Delta)$ in terms of the inverse of the PCA-weighted Euclidean distances defined by (4). For any specific neighbor k (k = 1, ..., K), such function is defined as follows

$$W(k,\Delta) = \Delta_k^{-1}\left(\sum_{k=1}^{K}\Delta_k^{-1}\right)^{-1} \qquad 0 < W(k,\Delta) < 1.0 \qquad \sum_{k=1}^{K} W(k,\Delta) = 1 \tag{6}$$

where k is a discrete random variable and $W(k, \Delta)$ represents its probability mass function. Note that $W(k, \Delta)$ decreases monotonically with the PCA-weighted Euclidean distance. The integration of $W(k, \Delta)$ over the K nearest neighbors works as the discrete cumulative distribution function (CDF). The random selection for resampling is based on a Monte Carlo simulation: in both the first and second disaggregation level, a uniform number $U(0,1)$ is generated. This number is used as the probabilistic metric in the CDF defined previously to select the nearest neighbor, which corresponds to one of the years

from the historical record. The discrete function in (6) can be also defined in terms of the inverse of the rank defined for the PCA-KNN (i.e., $k = 1, \ldots, K$), but such approach does not recognize the influence of the magnitude of Δ_k, and therefore, values of $W(k)$ will be always a constant probability for each performed simulation.

3.4.5 Step 5: Scaling and Concatenation

The resampled series x_{ip} (i.e., $\mathbf{X}$ in a vector form) is adjusted by means of a volume-dependent scaling factor. This procedure ensures conservation of mass, guarantees that no negative values are generated, and is not computationally intensive. A similar scheme based on this adjustment has been used in other works (Wood et al., 2002, 2004; Shaman et al., 2003; Salathé Jr et al., 2007; Lee et al., 2010). This procedure provides a simple alternative to the Gram-Schmidt orthonormalization procedure (e.g., Tarboton et al. 1998; Prairie et al., 2007; Bracken et al., 2010) that is prone to producing negative values. The disaggregated traces x^*_{ip} in the period p ($p = 1, \ldots, P$) are then

$$x^*_{i\,p} = V_p\, x_{ip} = \frac{y^*_p}{y_p} x_{i\,p} \qquad i = 1, \cdots, n \qquad (7)$$

where n is the vector size and the symbol * denotes the simulated (or downscaled) space. The scaling factor V_p in (7) is equal to the ratio between the seasonal (or monthly) volume to disaggregate, y^*_p, and the corresponding volume of the neighbor resampled, $y_p = x_{1p} + x_{2p} + \ldots + x_{np}$. Equation (7) defines a possible realization of the random process $\mathbf{X}^*$. Then a set of traces x^*_{ip} (e.g., each one associated to each one of the K nearest neighbors) constitutes an ensemble of $\mathbf{X}^*$ in a probabilistic sense. In the second disaggregation level, steps 1 - 5 are applied for each monthly volume. After each monthly

cycle in consecutive order (i.e., p = 1, 2, 3 and 4 for April, May, June and July, respectively), the disaggregated vector $\mathbf{X}^*$ at each period p is stored. $\mathbf{X}^*$ at p = 1, 2, 3 and 4 is a n = 30, 31, 30 and 31 -dimensional daily flow vector, respectively. After p = 4, all the $\mathbf{X}^*$s are concatenated (assembled) in the specified order to produce the disaggregated 122-dimensional daily flow vector.

3.5 Lag-1 Correlation (lag-1-R) Correction for MuNDi-S

Based on results presented in the model evaluation section (see section 4.5), the cascade procedure may lead to a significant underestimation of the lag-1 correlation between the flow of the last day of the previous month and that of the first day of the current month. In the spirit of the nonparametric (data-driven) strategy considered in this thesis, we present an algorithm based on nonparametric resampling to overcome such limitation.

Our approach for improving the preservation of the autocorrelation of flows between adjacent subperiods across consecutive periods in the aggregate space is remarkably different from that of traditional parametric models. Parametric approaches incorporate a term in the conditional $\mathbf{Y}$ (e.g., the underlying correlation needed to be preserved), or modify the linear stochastic model to include an additional linear dependency at the different levels of aggregation (Bras and Rodriguez-Iturbe, 1985). Instead, here we add an additional data point to the conditional $\mathbf{Y}$, which in our case corresponds to the last daily flow of the previous month. With this, the disaggregation framework preserves its parsimoniousness since the conditional PDF is still expressed in terms of nonparametric attributes, i.e., data-driven. The only parameter required in the

simulation process is the number of nearest neighbors K. Because our disaggregation framework is based on a resampling scheme of historical data, it preserves the sample serial dependence throughout the disaggregated period. In the procedure presented here, we redefine the resampling scheme described earlier by conditioning it on the last daily flow of the previous month. Suppose that the conditioning vector includes aggregate flows for the month p; then, using the same notation as presented earlier, the new conditional PDF is expressed as $\mathbf{f(X}_p|\mathbf{Y}_p, X_{np-1}\mathbf{)} = \mathbf{f(}X_{1p}, X_{2p}, \ldots, X_{np}\ |\mathbf{Y}_p, X_{np-1}\mathbf{)}$, where the new term X_{np-1} represents the value of the last daily flow of the previous month. The idea of conditioning the PDF on this additional observed (or estimated) flow data of the past period $p-1$ allows a strategy to connect the traces from the periods p and $p-1$, and thus, to improve the preservation of the observed serial dependence of flows across these periods. In particular, this is carried out first by computing new Euclidean distances in the disaggregated space between the last flow of the previous period and the potential K first flows of the ensemble of $\mathbf{X}^*$ (as defined in section 3.4.5) of the current period. Following, a modified discrete resampling function is applied over the nearest neighbors, which gives more weight to the smallest new Euclidean distances and less weight to the largest ones. One issue that arises at this point is related with the impact on the overall model performance for preserving historical features when permuting the nearest neighbors already sorted, i.e., when defining a new rank for the PCA-KNN classification described in section 3.4.3. Such results are presented in section 4.5.

The procedure presented here is understood as optional if the modeler wishes to improve the preservation of the serial dependence between the components of $\mathbf{X}^*$s across months with insignificant impacts on preserving other statistical features. Because this

procedure to improve the lag-1-R correlation can be applied at each month in the second disaggregation level (or where it is necessary), from here on the subscript (*i, p*) on the time series x_{ip} is replaced by the subscripts LDPM and FDCM when (*i, p*) refers to the "last day of the previous month" and the "first day of the current month", respectively. The steps of the lag-1-R adjustment algorithm are as follows:

Step 1. At the first month (*p= ini*) of the second disaggregation level, a daily sequence $\mathbf{X}^*$ (i.e., $x^*_{i(ini)}$; *i=1, ..., n*) is simulated with MuNDi-S and stored.

Step 2. At the next month (*p= ini+1*), *K* nearest neighbors are identified and sorted from 1 to *K* according to steps 1-3 of the MuNDi-S algorithm. This step selects and ranks the historical years with the most similar hydroclimatic behavior to that of the query point.

Step 3. An ensemble of $\mathbf{X}^*$ (i.e., *K* daily traces $x^*_{i(ini+1)}$; *i=1, ..., n'*) is generated, as described in step 5 of MuNDi-S. This step defines a potential set of traces that can be used as disaggregated trace at *p= ini+1*. At this point, no stochastic selection over the nearest neighbors has been made at *p= ini+1*.

Step 4. New Euclidean distances, now in the disaggregated space, are then computed between the flow of LDPM, $x^*_{(LDPM)}$ (i.e., at *p= ini*), and the *K* flows of FDCM, $x^*_{(FDCM)}$ (i.e., at *p= ini+1*), as follows

$$\Delta^*_k = \left| x^*_{(LDPM)} - x^*_{k(FDCM)} \right| \qquad for\, k = 1,2,\ldots,K \tag{8}$$

These new Euclidean distances are used in a modified resampling function defined bellow.

Step 5. The rank of PCA-KNN at $p = ini+1$ and their associated $x^*_{i(ini+1)}$ are reorganized, by sorting values of Δ^*_k in ascending order. As a consequence, a new ranking from 1 to K is assigned (1 for the smallest Δ^*_k and K for the largest Δ^*_k).

Step 6. The neighbor is now randomly selected using the Monte Carlo simulation described in step 4 of MuNDi-S, but using a modified discrete resampling function $\Omega(k, x^*)$ (evaluated in $x^*_{(FDCM)}$) that decreases as the distance from $x^*_{(LDPM)}$ increases

$$\Omega(k, x^*) = \frac{1}{\left|x^*_{(LDPM)} - x^*_{k(FDCM)}\right|} \left(\frac{1}{\sum_{k=1}^{K} \frac{1}{\left|x^*_{(LDPM)} - x^*_{k(FDCM)}\right|}} \right)$$

$$0 < \Omega(k, x^*) < 1.0 \qquad (9)$$

$$\sum_{k=1}^{K} \Omega(k, x^*) = 1$$

Step 7. The disaggregated vector $\mathbf{X}^*$ from step 5 corresponding to the randomly chosen neighbor from step 6 is selected and stored.

Step 8. Continuing with the stepwise scheme for the rest of months, the last disaggregated month (into daily) is set up now as the first month ($p = ini$), and the steps 2-8 are repeated consecutively.

For the period of study here, step 1 is only applied in April, and steps 2-8 are repeated for May, June and July, respectively. Note that when using MuNDi-S, only one synthetic daily trace is generated at each month; but with the required ensemble when incorporating the Lag-1-R correction algorithm, our modeling framework becomes slightly computationally more expensive. This nonlinear adjustment approach can also be applied at other time scales (e.g., for annual to monthly disaggregation, for preserving the

correlation between December of the previous year and January of the current year), and offers a simple and parsimonious alternative to existing parametric correction procedures (e.g., Koutsoyiannis and Manetas, 1996).

3.6 Model Evaluation

Both the fitting performance and the ability to preserve historical features of the MuNDi-S model are assessed with data from the Payette River, Idaho. A brief hydrologic description and the streamflow data source of the study region are provided in section 2.1. The fitting performance evaluation includes also information of large-scale climatic patterns in the Pacific Ocean, largely described in section 2.2. The observational record considered in this analysis is from the period 1950–2007.

3.6.1 Fitting Performance and Variable Selection

A leave-one-out cross-validation (LOO-CV) combinatorial approach (referred as to "LOO-CV combinatorial") was implemented in order to assess the model performance in terms of predicting observed flow sequences. The combinatorial scheme involves different levels of combinations among periods and variables that comprise **Y**. Nine periods for each averaged climatic index are analyzed. As stated in the MuNDi-S description (section 3.4), each index (PDO, SOI and NPGO) is averaged over a 4-month interval leading the spring runoff season. The first period leading the spring runoff starts in May of the previous year and the last one starts in January of the current year. The maximum number of possible combinations in **Y** regarding the nine climatic periods for the three climatic indices is $9^3 = 729$. Although a random selection of one of the

neighbors is considered in the bootstrap approach (see section 3.4.4), we forced the MuNDi-S algorithm to resample the same k^{th} nearest neighbor at each disaggregation level for further analyses of the fitting performance. Although the possible number of combinations between variables of the vector **Y** is $2^m - 1 = 15$, selected results from those subsets that include the streamflow variable are presented. Eight combinations were analyzed: (Y_1), (Y_1, Y_2, Y_3, Y_4), (Y_1, Y_2, Y_3), (Y_1, Y_2), (Y_1, Y_3), (Y_1, Y_2, Y_4), (Y_1, Y_3, Y_4) and (Y_1, Y_4). The subset (Y_1) implies that only hydrological information is considered. In the LOO-CV scheme, N years are available for developing the model. Here, we use N-1 years for training MuNDi-S leaving the j^{th} year out (j=1, ..., N), and the fitted model is validated in that j^{th} year. This means that we use observational information to construct the j^{th} query point and discard its corresponding historical **Y** in defining the **f(X,Y)** data space. The LOO-CV combinatorial evaluation tries to determine not only which season/variable subsets produce the best performance, but also to compare performances between different modeling procedures (e.g., like daily downscaling using hydrological information alone and/or without the stepwise procedure). We compared graphically and numerically a single disaggregated trace with the observed one of the year j^{th} (j=1, ..., N) that was dropped from the data. To numerically assess the model performance, the mean square error (MSE) was computed as

$$MSE(x,p,j) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ip}^{*(j)} - x_{ip}^{(j)}\right)^2 \qquad \begin{array}{l} for\ j = 1,2,\ldots,N \\ p = 1,2,\ldots,P \end{array} \tag{10}$$

with x^*_{ip} and x_{ip} as defined by (7). We used the total MSE (TMSE) as the performance metric of the synthetic daily traces from April through July at each season/variable subset and for each neighbor (from $k = 1$ to $k = K$)

$$TMSE(x) = \frac{1}{\sum_{p=1}^{P} n_p} \sum_{j=1}^{N} \sum_{p=1}^{P} \sum_{i=1}^{n} \left(x_{ip}^{*(j)} - x_{ip}^{(j)} \right)^2 \qquad (11)$$

The value of N used in (10) and (11) is 58. Both $\mathbf{X}$ and $\mathbf{X}^*$ at p = 1, 2, 3 and 4 are n = 30, 31, 30 and 31 -dimensional daily flow vectors, respectively. Other metrics like the average MSE (AMSE) or the average root MSE (ARMSE) instead of the TMSE, the average R, or a bias-based measure could be tested also. For the selection of variables we evaluated the best combination in terms of the minimization of the TMSE. For evaluating the effect of the cascade scheme in the disaggregation framework, the stepwise algorithm was disconnected from MuNDi-S, i.e., the spring natural runoff volumes were disaggregated directly into daily values.

3.6.2 Preservation of Statistical Properties

In general terms, we simulate first an aggregate flow space with a suitable approach, and then we generate stochastically in a finer time scale with the MuNDi-S model. The goal of this simulation is to assess the reproduction of statistical attributes from the historical record in the disaggregated space. Parametric and nonparametric schemes can be used for generating the aggregate flow space. Salas and Lee (2009) provide an interesting review of nonparametric-based generation models, describe their main strengths and drawbacks, and propose a new framework for overcoming their limitations. A univariate AR(1) model was implemented here for generating seasonal

(i.e., aggregate) flows (here, 200 sequences of spring flows, each one with length = N). The parameter estimation of AR(1) was carried out using the method of moments (Salas et al., 1980) and based on the available sample size, 1950-2007 (N = 58 years). Such model was fitted directly to the original data space because the observational spring flow series appears not skewed (a skewness test of normality (Salas et al., 1980) was applied). However, a Gaussian transformation can be considered previous to the parameter estimation. We implemented the MuNDi-S approach to disaggregate the synthetic spring flows into monthly values. Following, we simulated daily runoff traces with MuNDi-S from the synthetic monthly volumes. Only hydrological information is considered in the aggregate vector. A similar and consistent procedure also applies when disaggregating directly from seasonal to daily values.

We evaluated the reproduction of several sample moments and statistics (e.g., estimates of the mean, variance, skew, lag-1 serial correlation, maximum and minimum). To test the ability of MuNDi-S to capture the historical distributional properties, comparisons between synthetic and observed marginal PDFs of both monthly flows in the first disaggregation level and daily flows in the second disaggregation level were carried out by means of univariate KDE (Simonoff, 1996). Box-and-whisker vertical diagrams were used in all the assessed statistical features to quantify variability (200 realizations at each box). The line dividing the box represents the median. The top (75% quartile) and bottom (25% quartile) of the box encompass the interquartile range (IQR), and whiskers span the 5% and 95% quartiles. The span of the IQR and whiskers represents the underlying variability. Circles within the boxes correspond to historical statistics. Outliers from the simulations were omitted in the boxplots. In what follows, the

performance of the MuNDi-S model in reproducing statistical attributes will be referred to as *very good* if the historical sample statistic coincides with the median, *good* if it falls inside the IQR, *acceptable* if it falls inside the whiskers frame and *poor* if it falls outside the box-and-whisker. The algorithm described in section 3.5 for fixing the autocorrelation structure between daily flows across months of the MuNDi-S model was also tested with this evaluation framework.

4 Results

4.1 Lagged Teleconnections with SST and SLP Fields

The spatial and lagged cross-correlation variability between SST and SLP fields in the Pacific Ocean and spring streamflows in the Payette River is displayed in Figure 5 and Figure 6, respectively. The images were processed at the Earth System Research Laboratory with data from NOAA. Pearson's *R* equal to or above/below ±0.255 (df = 57), respectively, are statistically significant at the 95% confidence level. Although correlation maps were generated with climatic fields leading the spring runoff up to ten months, only results from the preceding fall and winter are presented here because the observed SST-based correlations are higher in these seasons.

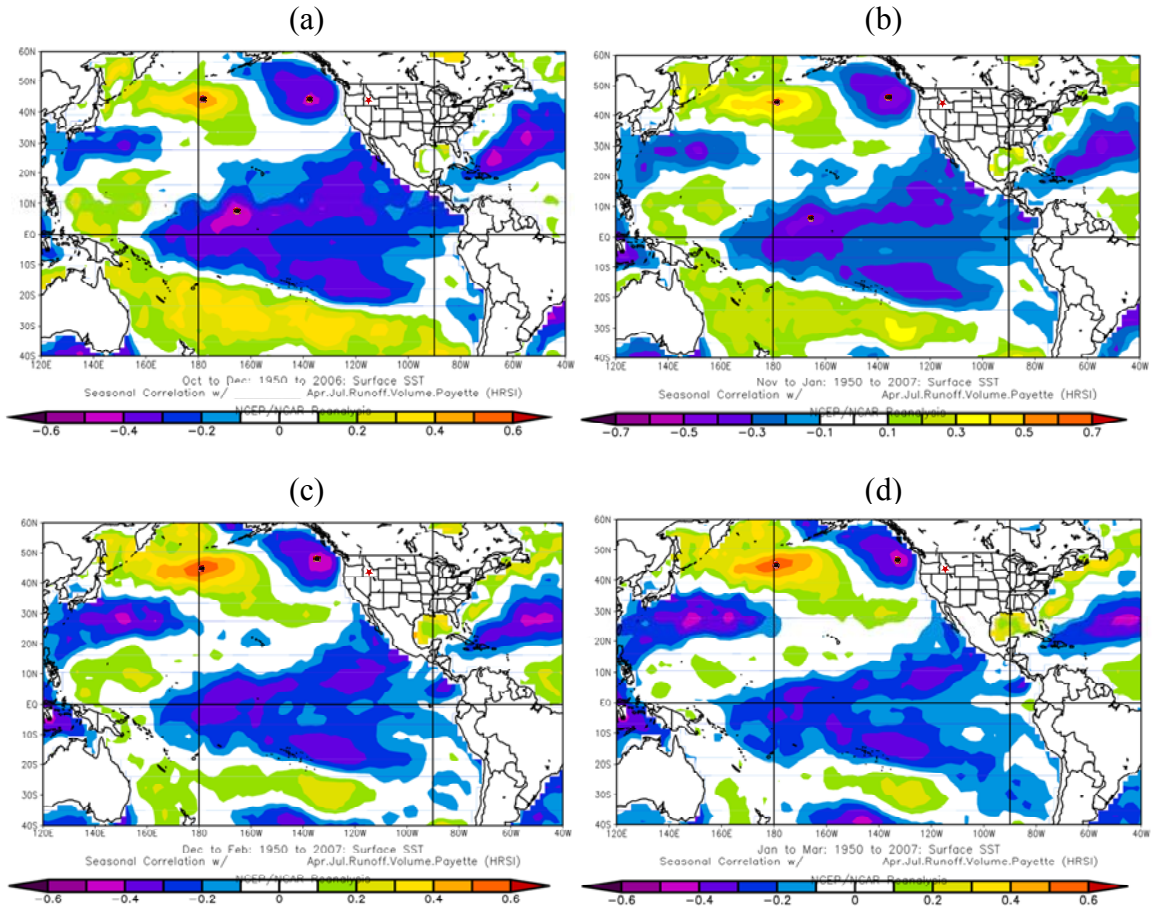

Figure 5. Spatial variability of cross-correlation between spring runoff volumes in the Payette River Basin, Idaho, and preceding SSTs averaged over: (a) October–December; (b) November-January; (c) December-February; (d) January-March. Pearson's *R* equal to or above/bellow ±0.255 are 95% statistically significant. The regions with the highest positive and negative correlations in the Pacific Ocean are depicted with black circles and a red star points the location of the Payette River. (Image processed by the NOAA-ESRL Physical Sciences Division, Boulder, Colorado, USA).

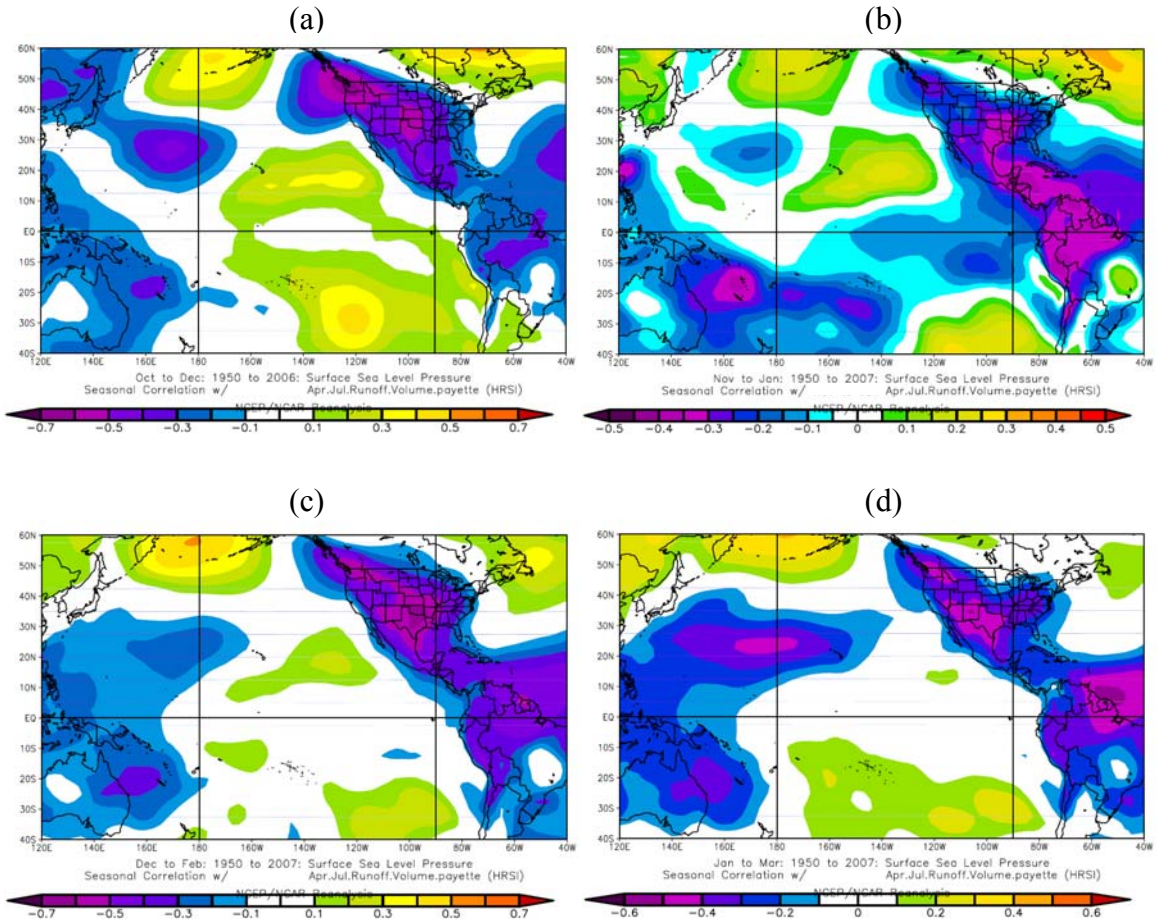

Figure 6. Same as Figure 5 but for SLP fields. Image processed by the NOAA-ESRL Physical Sciences Division, Boulder, Colorado, USA.

It can be seen in Figure 5 that as the period of analysis in the north Pacific moves from fall to winter both the correlation level and sea surface areas with the highest *R* values increase. In the tropics, on the other hand, the correlations are stronger in the fall than in the winter. In general terms, these patterns could be the result not only of geographical reasons, (i.e., distances from the tropics to Idaho are greater than those from the northern Pacific ocean and, consequently, atmospheric perturbations induced from

anomalously warm or cold SSTs in the tropics take more time to impact the North Pacific U.S. atmospheric circulation than those from the northern Pacific ocean) but also of "atmospheric bridges" in which part of ENSO-related phenomena leads and influences the ocean-atmosphere oscillations in the North (Trenberth and Hurrell, 1994; Newman et al., 2003; Di Lorenzo et al., 2009), thus, causing a delay in the North Pacific oceanic response and the associated hydroclimatic impact in the continent. Payette River flows and SSTs in the Tropical Pacific Ocean display negatively correlated activity, although some western regions in this oceanic frame (e.g., north and off the coast of Papua New Guinea) exhibit positive links. Significant positive and negative associations are observed in the northwestern and northeastern Pacific Ocean, respectively. The highest levels of correlation are located in regions of the North Pacific (between 40°N and 50°N) and in the tropical Pacific (between equator and 10°N). In the former regions the highest positive correlations are approximately located at longitude 178°W and those areas with the highest negative values are located in the region off the coast of the states of Washington and Oregon (Figure 5). Regions in the tropical zone with the highest negative correlation are located in the central Pacific, between 160°W and 165°W (i.e., corresponding to the region Niño 3.4 which is defined for 5°N-5°S, 120°W-170°W). The highest cross-correlation in the Northern Hemisphere is observed in the periods December-February and January-March (Figure 5c and 5d, respectively), with positive values close to 0.6 in western regions and negative values close to 0.5 just off of the Washington and Oregon coasts, which suggests that such SST anomalies explain up to 60% of the spring flow variability in the study zone. Similarly but to a lesser degree, the most significant period for the central tropical Pacific is October-December (Figure 5a)

with a negative correlation close to 0.5. It is also observed that fall and winter SSTs in the northern hemisphere always have the highest correlation with respect to other regions. For the period January – March, Figure 5d shows that the strength of association in the central tropical Pacific decreased with respect to preceding periods but still with significant cross-correlation levels in the north.

Taking into account the definition of the SOI provided earlier in this thesis, an additional SLP-based correlation analysis in the Tropical Pacific Ocean was performed in terms of the absolute difference of cross-correlation values between Darwin and Tahiti (ADCDT), and not only as a function of the single values at each location. In the South Pacific Ocean basin, east of Australia, the SST-runoff correlation structure from Figure 6 exhibits a negative association, and a vast area centered at approximately 20° S and 160° E shows always significant and higher correlation values in this region ($-0.4 < R < -0.3$). The SLP-based fields at the Darwin location show also negative values but less correlated. However, at Tahiti, neutral and slight tendency to weak positive relationships ($R < 0.2$) can be seen. In average, the ADCDT remains moderate during the period of analysis (ADCDT ≈ 0.3), as shown in Figures 6a – 6d. However, we find that in the preceding summer (results not shown), the ADCDT indicates a stronger relation (ADCDT ≈ 0.6) than this from fall and winter due to higher and statistically significant correlations over the central tropical Pacific. This is consistent with results presented later from SOI-based correlogram functions. Also, a longer memory is apparent from SST-based teleconnections when compared with those from SLPs. This is possibly due to the higher heat capacity of the ocean, which plays a more important role than the atmosphere in the storage of heat and likely provides a longer persistence for the SST-runoff

relationship. In general terms, the significant correlations found in SST and SLP fields are consistent with those regions in the Pacific Ocean where PDO, NPGO and ENSO fluctuations are defined.

4.2 Lagged Teleconnections with PDO, NPGO, and ENSO Indices

Figures 7–9 show the cross-correlograms between spring runoff and preceding indices of PDO, NPGO and ENSO/SOI, respectively. Figure 7a shows that for lag-times larger than zero and for all the assessed periods the correlation is poor and not significant (with very few exceptions from lag-3 to lag-5). Such results indicate no interannual persistence in the PDO-runoff variations. This is in contrast with the significant correlation and memory observed at the seasonal timescale (Figure 7b) up to six lags. These analyses suggest that predictors involving PDO and spring runoff variables in Payette should be preferably selected with averaged indices of PDO leading the spring runoff by less than or equal to eight months (e.g., up to the preceding period August-October), and not by a number of years. For the NPGO case (in contrast to the PDO), this analysis shows positive and significant associations with runoff for most of the preceding periods in lags from three up to seven years (Figure 8a), and accounting for nearly 40% of the streamflow variance. A similar behavior is also observed for lag-10 through lag-13 but with negative correlations. These correlation patterns are likely associated to the high low-frequency found for the NPGO (see spectral analysis, section 4.3). Such lagged teleconnections indicate that potentially there may be some long-lead predictive ability of the NPGO in an annual time scale. Similarly but less pronounced, the cross-correlation runoff-SOI is significant mainly at lag-3 and lag-12 (Figure 9a) for all the periods

evaluated, with an average of explained runoff variability of about 30% and 36%, respectively.

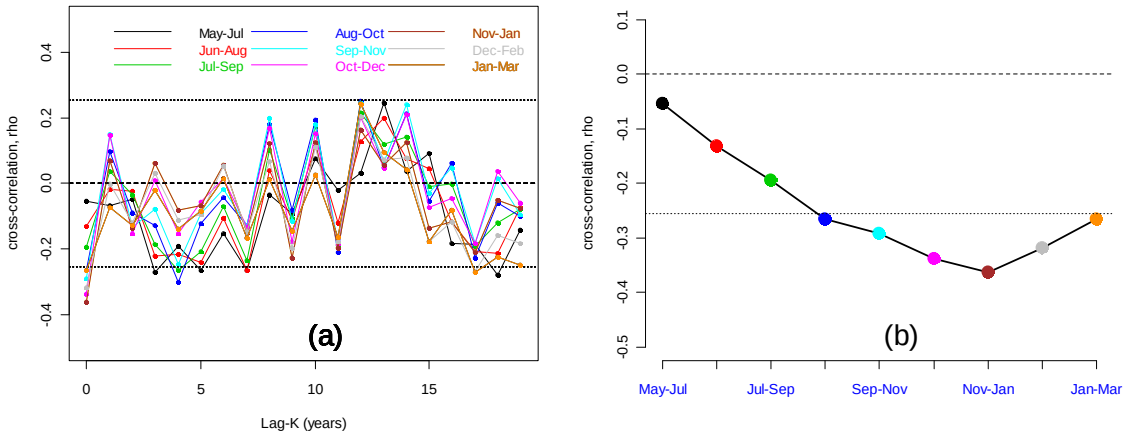

Figure 7. (a) Lag-K cross-correlation functions between spring runoff volumes at Payette River at Horseshoe Bend, Idaho, and 3-month average PDO index for several preceding periods. (b) Variation of Lag-0 cross-correlation between PDO index and spring runoff volumes at Payette River. Pearson's *R* values above 0.26 and below –0.26 are significant at the 95% level or greater.

Figure 7b shows negative and statistically significant correlations for the PDO-runoff variations at lag-0, and basically spanning the fall and winter and part of the summer. Such negative relations indicate generally lower flows during the positive phase of PDO and vice versa (see Loess smoothing in Figure 10). The period with the strongest significant lag-0 cross-correlation is November-January (R = -0.36). The NPGO (Figure 8b) also presents negative associations as PDO at lag-0, but exhibiting a tendency to more uniform and weaker correlations than the PDO, and only significant in the winter.

These results show consistency with the SST-based negative relationships observed for the same seasons in the region off the coast of Washington and Oregon (Figure 5), where the Alaskan gyre, a wind- and buoyancy-forced circulation regime located in the northwest Pacific and the Gulf of Alaska, plays an important role in defining the PDO and NPGO patterns (Di Lorenzo et al., 2008). The season with the higher lag-0 cross-correlation in the runoff-NPGO variations is January-March (R = -0.28). The weaker links of the NPGO when compared with those of the PDO in the first lags of the seasonal timescale can be explained partially because the former represents the second mode of variability of the North Pacific SSH field (the PC2 of SSTs is also an expression of the NPGO), and thus, it accounts for less variance than the PDO. However, the more prominent low-frequency fluctuations found in the NPGO with respect to the PDO, as mentioned later, and the related atmospheric mechanisms that drive that behavior may play an additional role in the important runoff variability explained by this oceanic circulation in annual timescales (e.g., $0.35 < R < 0.45$ for lag-5, Figure 8a).

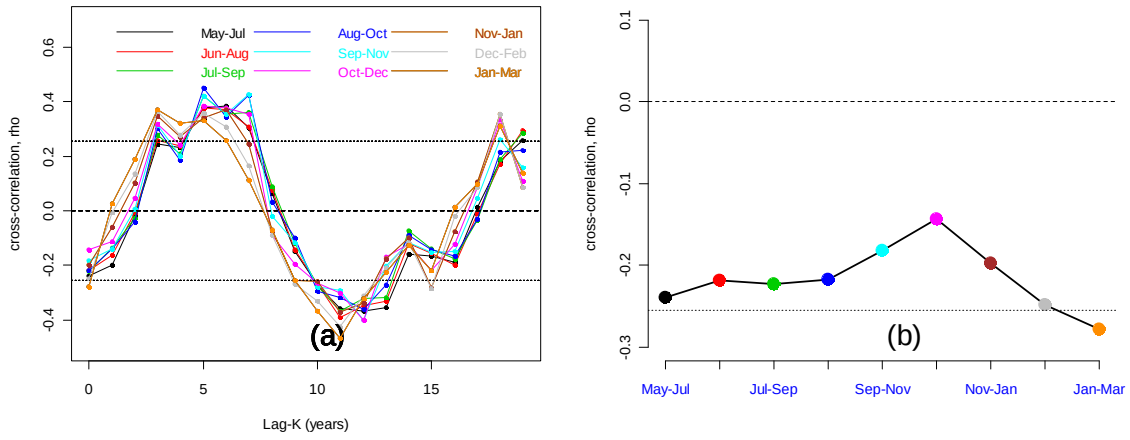

Figure 8. Same as Figure 7 but for NPGO index.

Significant and positive associations for the SOI-runoff at lag-0 are observed mainly in summer, with the highest value in June-August (R = 0.34; Figure 9b). This is consistent and supports previous analysis carried out in the western tropical Pacific with the ADCDT. Note that the climatic indices used in this study exhibit less cross-covariance with the spring runoff than some specific regional climate fluctuations in the Pacific basin. For example, December-February SSTs in the region 178°W-182°W and 40°N-50°N explains around 60% of the spring runoff variability (Figure 5c), whereas the PDO index for the same period explains around 33% of the spring runoff variability (Figure 7b). This is because such indices explain partially the variability associated to the climate fields where they are defined. In the case of the PDO index, for example, it accounts for 34% of the SST field variability in the North Pacific Ocean (Di Lorenzo et al., 2008), which corresponds to the variance associated with the PC1 of SSTs in that region. Bearing this in mind, more specific regional indices in the ocean-atmosphere system associated with the study basin should be developed and explored in future research to be used as inputs in the disaggregation framework.

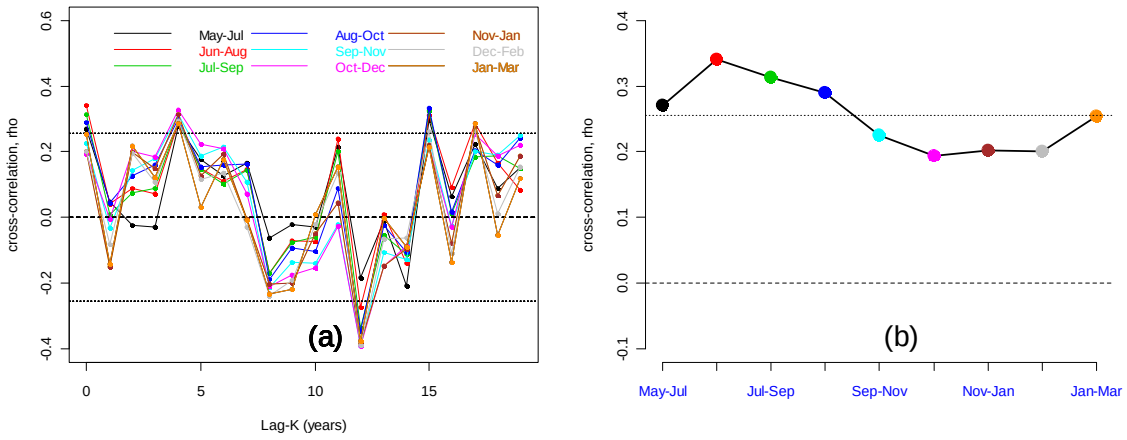

Figure 9. Same as Figure 7 but for SOI.

Climate indices with the highest absolute values of Lag-0 cross-correlations were statistically linked between them, as shown in the scatterplots in Figure 10. Pairwise relationships (lower panel) described by the conditional expectation (smoothed curves) estimated with Loess exhibit slight nonlinearities in most cases, which may be related to differences in their predominant frequencies of oscillation. Regarding the variation of the runoff-index smoothing (e.g., $E[Vol_{Apr.Jul}|PDO_{Nov.Jan}]$ or $E[Vol_{Apr.Jul}|SOI_{Jun.Aug}]$) in the same figure, it is noted that high and low streamflow years exhibit a strong linear relationship with PDO and SOI, but in normal years (standardized flows close or equal to zero) the locally-weighted relationship is observed to be flattened, i.e., PDO or SOI alone fails to explain the streamflow variability. It is thought that such nonlinearities may be associated to the modulating effect from phase interactions of these climate fluctuations over the hydrologic response. The effects of PDO phase on SOI-streamflow relationships in the Pacific Northwest U.S. have been previously studied (Koch and Fisher, 2000;

Harshburger et al., 2002; Barton and Ramirez, 2004), and they indicate that runoff responses to El Niño events during the positive phase of PDO are stronger and similarly when La Niña events occur during the negative PDO, whereas events with equal sign in these two fluctuations tend to weaken their effects on the streamflow variability.

NPGO (Nov-Jan) is intentionally added to Figure 10 in order to make inferences of independence with the PDO for the same season. Note that Pearson's *R* values in the correlation matrix (upper panel) appear as absolute values, but the sign of association between spring runoff and PDO/NPGO indices is negative, as can be seen in the smoothed curves in the scatterplots. PDO (Nov-Jan) and NPGO (Nov-Jan) are not statistically correlated, as conceptually defined, although the small observed correlation ($R = 0.13$) is explained because the PDO used in this analysis is derived from the SST field and not from the SSH anomalies.

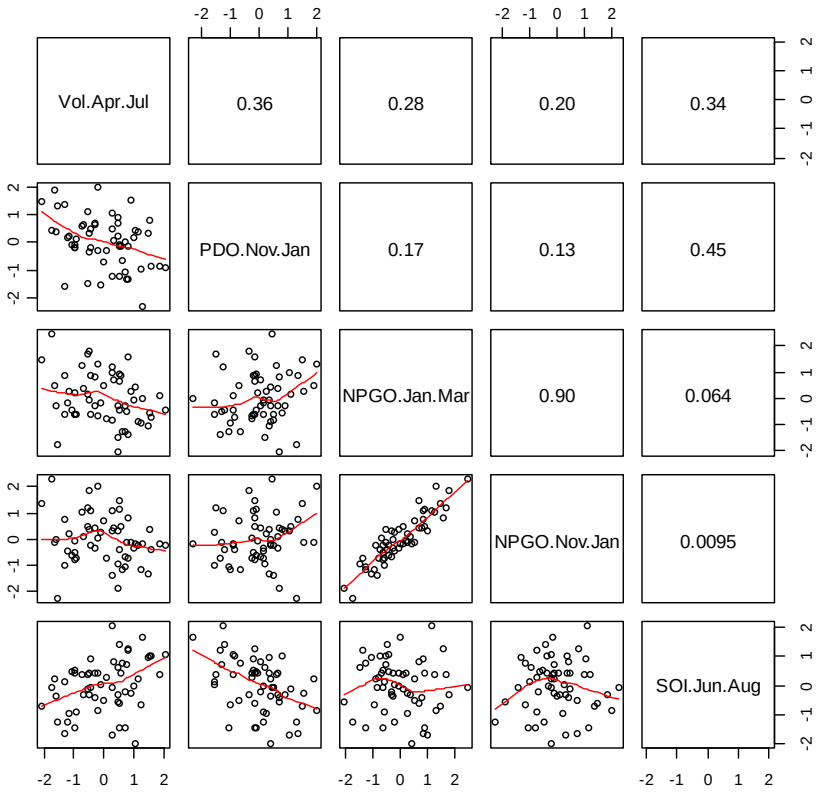

Figure 10. Lower panel: Pairwise scatterplots between spring runoff volumes (Vol. Apr-Jul) at Payette River near Horseshoe Bend, Idaho, and preceding PDO (Nov-Jan), NPGO (Jan-Mar), NPGO (Nov-Jan), and SOI (Jun-Aug) indices; Upper panel: Absolute values of Pearson correlations. The lines in scatterplots smoothed with Loess.

Attention is focused now on the statistical links between the NPGO and SOI indices. Based on results shown in Figure 10, the NPGO (Jan-Mar) and preceding SOI (Jun-Aug) are uncorrelated (R = 0.06), and NPGO (Nov-Jan) and preceding SOI (Jun-Aug) are independent (R ≈ 0). Other cross-covariance results (not shown) considering a range of seasonal series indicate that the correlation between these two patterns are not

significant statistically, specifically the correlation between the fall-winter NPGO and the preceding spring-summer SOI. In contrast, PDO (Nov-Jan) and SOI (Jun-Aug) are significantly correlated, and appear to exhibit a significant negative association (R = -0.45). Similarly, the correlation between PDO and SOI for other periods (figures not shown) is negative and statistically significant (R < -0.26). The previous analyses suggest that ENSO and NPGO should be explored as an alternative to ENSO and PDO in operational hydrology over the western U.S.

4.3 Spectral Analysis of PDO, NPGO, and ENSO Indices

Figure 11 shows the spectrum for the climatic indices considered in this manuscript. The value of the bandwidth that appears in the plots refers to the width of the frequency band used in smoothing the periodogram (further details for the bandwidth computation can be found in Shumway and Stoffer (2006) and references therein). Although the abscissa spans up to j/n = 0.5, we show values of j/n from 0 to 0.06, where the main frequencies are observed. The magnitude of the spectrum for Fourier frequencies larger than 0.1 is minimum or tends to zero. In average, the dominant oscillation rates estimated for the PDO, NPGO, and SOI are 1/648 (54 years), 1/144 (12 years), and 1/60 (5 years) cycle/month, respectively. Two other modes of oscillation for the SOI in the interannual timescale are also observed around j/n = 0.033 cycle/month (30 months) and 0.022 cycle/month (45 months), and account for a significant amount of its variance. The fourth SOI mode of periodicity, but in the decadal time scale, is observed around j/n = 0.0069 cycle/month (144 months).

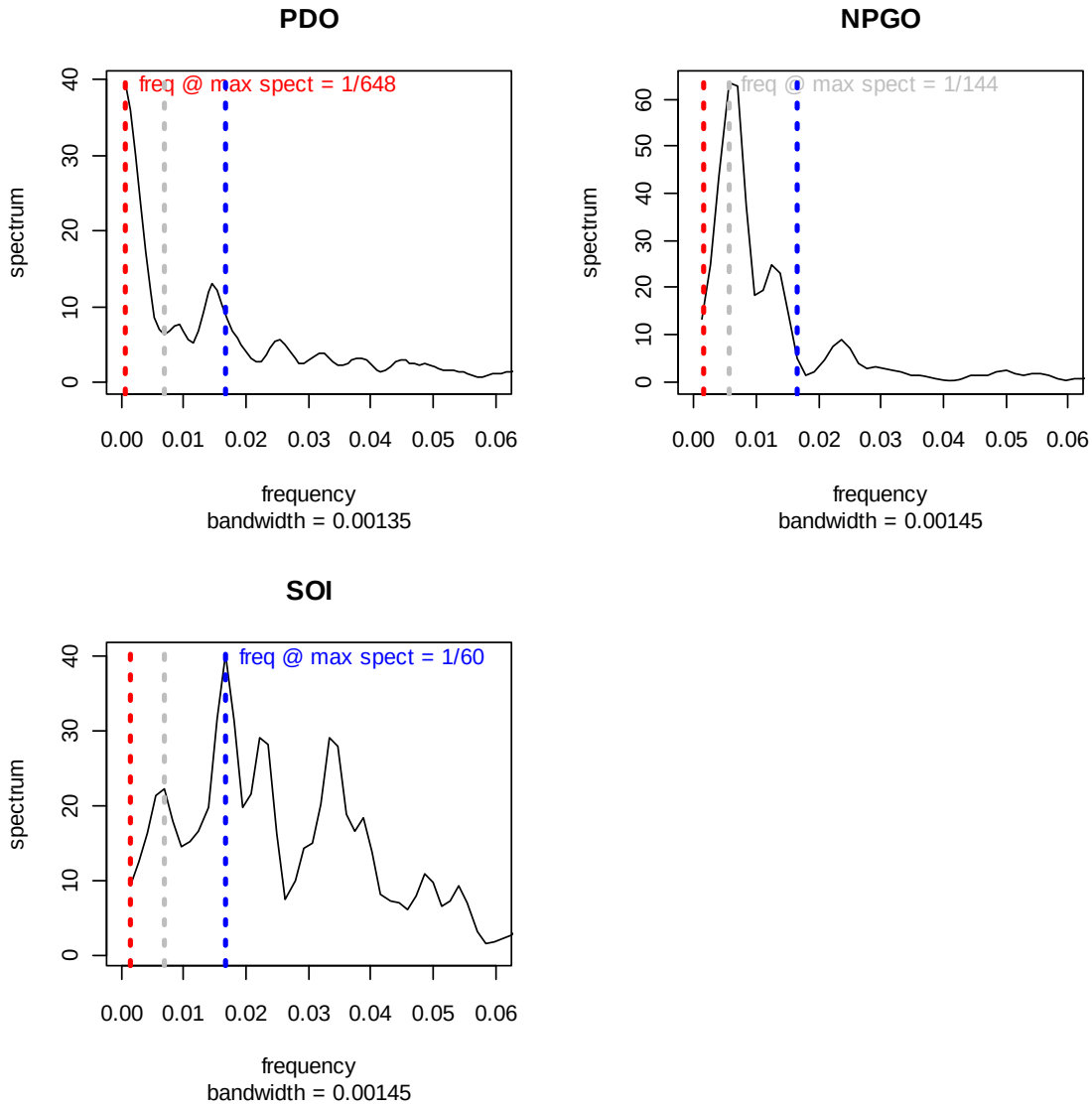

Figure 11. Smoothed spectrum for PDO, NPGO, and SOI signals. The spectral estimates smoothed with the modified Daniell kernel. The dashed vertical line depicts the predominant oscillation frequency (red for PDO, grey for NPGO, and blue for SOI). Fourier frequency units are given in cycle/month. A coherence between SOI and PDO around $j/n = 0.0166$ (~5 years) and between SOI and NPGO in $j/n = 0.007$ (~12 years) is observed.

The interdecadal dynamics estimated with the PDO spectrum in Figure 11 is easily visualized in the smoothed series using the moving average function plotted in Figure 3, and is consistent with Mantua et al. (1997), who characterize PDO events (positive or negative phases) with a persistence over 20 to 30 years. A second mode of periodicity but in the annual timescale is also observed for the PDO, and very closely tracks the main periodicity of the SOI. Such coherence between these two patterns in the annual timescale supports results of their statistical association described in the previous section, although the relationship between ENSO and the PDO extends to longer timescales in the power spectra domain, as noted by Newman et al. (2003).

Our spectral analysis indicates that the NPGO exhibits more prominent low-frequency oscillations than the PDO. The spectral peak found for the NPGO at a frequency of about one cycle every twelve years (Figure 11) is consistent with the periodic variation plotted for this pattern in Figure 4. Although in the previous section we found statistical independence between NPGO and SOI at seasonal time scales (0-6 months), our results from the spectral analysis suggest that there is a link between these two patterns, where the SOI signal appears excited by the low-frequency oscillation mode of the NPGO. As observed in Figure 11, the fourth SOI frequency tracks exactly the main periodicity of the NPGO. Recent theories suggest that the NPO and NPGO are connected to ENSO and act as the drivers of SST anomalies in the central tropical Pacific (Vimont et al., 2009; Di Lorenzo et al., 2009). The spectral link seen in Figure 11 between NPGO and SOI is consistent with results of Di Lorenzo et al. (2009), who found that the monthly time series (1950-2006) of the first principal component of ENSO (as defined therein) is significantly correlated with those of the NPGO index when the latter leads the first by

approximately one year. They hypothesize that boreal spring variability of the NPO, integrated into the oceanic NPGO pattern, initiates the ENSO expression that peaks in the following winter, which in turn excites variability in the AL that is passed by the ocean into the PDO pattern.

4.4 Fitting Performance

Four modeling alternatives are analysed: (1) MuNDi-S with hydroclimatic data in **Y**, (2) MuNDi-S with only flows in **Y**, (3) alternative 1 without stepwise disaggregation, and (4) alternative 2 without stepwise disaggregation. The ability of the modeling procedures to simulate observed flows is based on comparisons between minimum TMSE values, and therefore these metrics are used to determine which alternative produces the best performance. The model ability to reproduce statistics of historical data is tested in the next section. Figures 12 - 14 present the MuNDi-S performance for alternatives 1 and 2 for selected variable subsets and for the first six neighbors. Similarly, Figures 15 – 17 show selected results of performance for alternatives 3 and 4. Although such figures do not represent time series, they were plotted conveniently in that way as a valuable tool for visualizing and discussing results. Note that the TMSE for alternatives 2 and 4 (depicted with red points in their respective figures) is constant throughout the lagged climatic periods evaluated at each nearest neighbor. This, because the vector **Y** considers aggregate flows only in the same season/month where the disaggregation is carried out (i.e., $\mathbf{f(X}_p|\mathbf{Y}_p\mathbf{)}$).

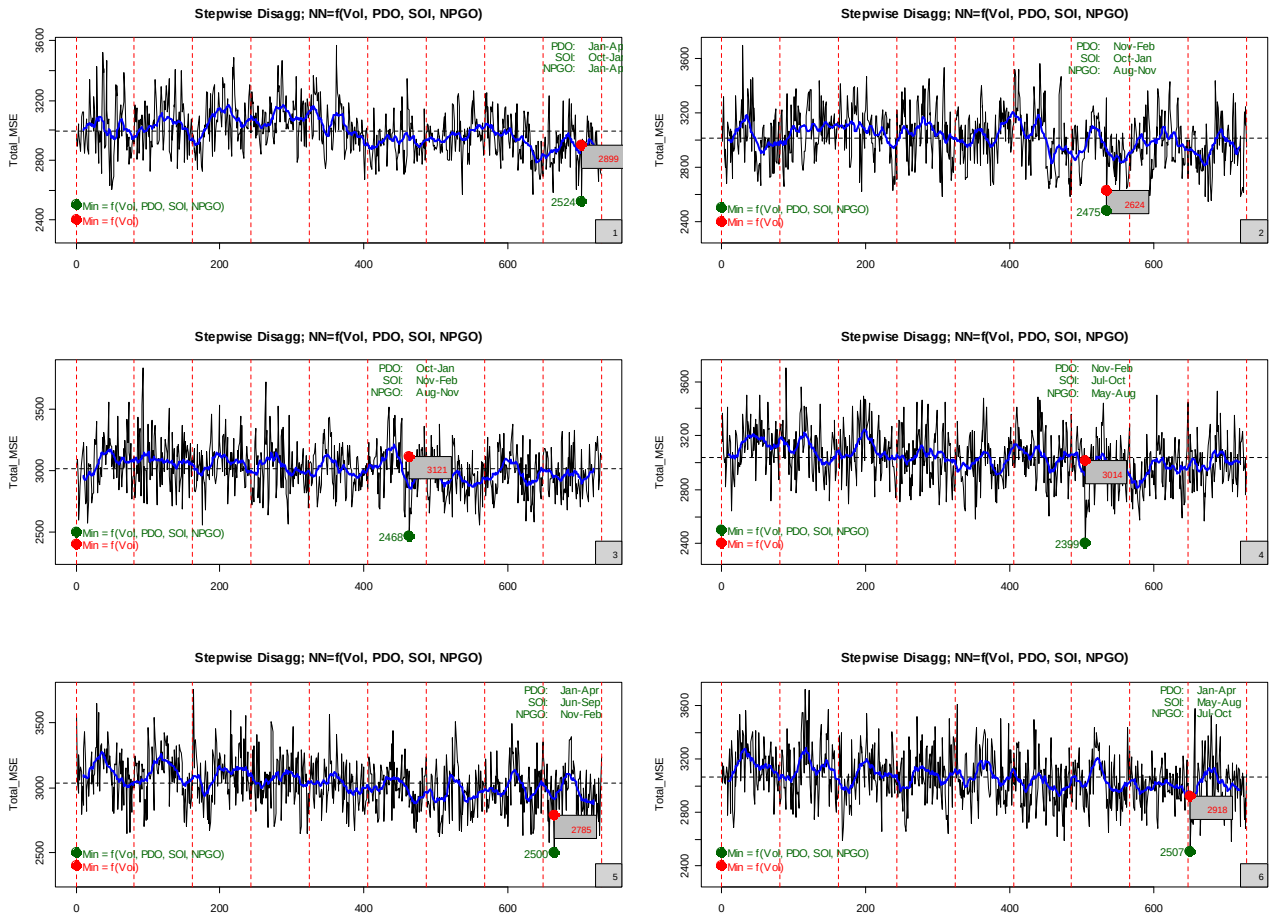

Figure 12. MuNDi-S performance [TMSE in units of $(10^6 m^3/d)^2$] conditional on **Y** = (*Vol, PDO, SOI, NPGO*) for spring to daily disaggregation of flows at Payette River using LOO-CV combinatorial (black line). Each integer in the abscissa represents one possible combination of 4-month periods from climatic indices in the conditional **Y** (number of combinations = 729). The red points depict the model performance conditional on flows alone (alternative 2). The k^{th} NN evaluated is printed in the right lower corner of each plot. The horizontal dashed line and the blue line represent the mean TMSE and a smoothed TMSE with a moving average function, respectively. The 4-month periods subset that produces the best disaggregation (shown in front of each index) corresponds to the combination defined by the minimum TMSE (green points).

Figure 12 shows that using flows and the three climatic indices in the conditional vector improves significantly the fitting performance at each nearest neighbor when compared with disaggregating using flows alone. Other combinations from alternative 1 like those presented in Figures 13 and 14 exhibit similar improvements over using flows alone, but their performance is not as good as that of alternative 1 shown in Figure 12.

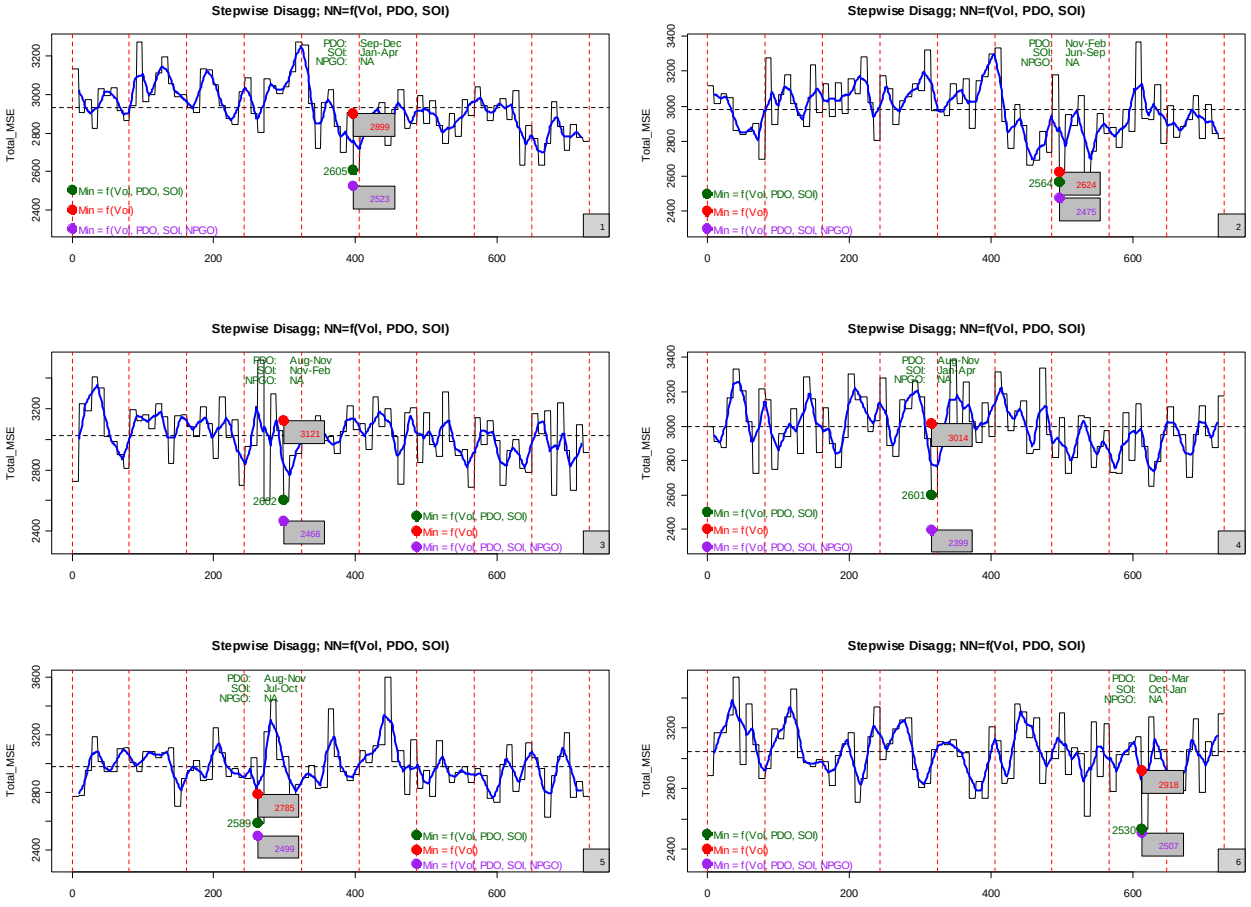

Figure 13. Same as Figure 12 but for **Y** = (*Vol, PDO, SOI*). Number of combinations of seasonal periods = 81, instead 729. Results of minimum TMSE for **Y** = (*Vol, PDO, SOI, NPGO*) are also included (purple point).

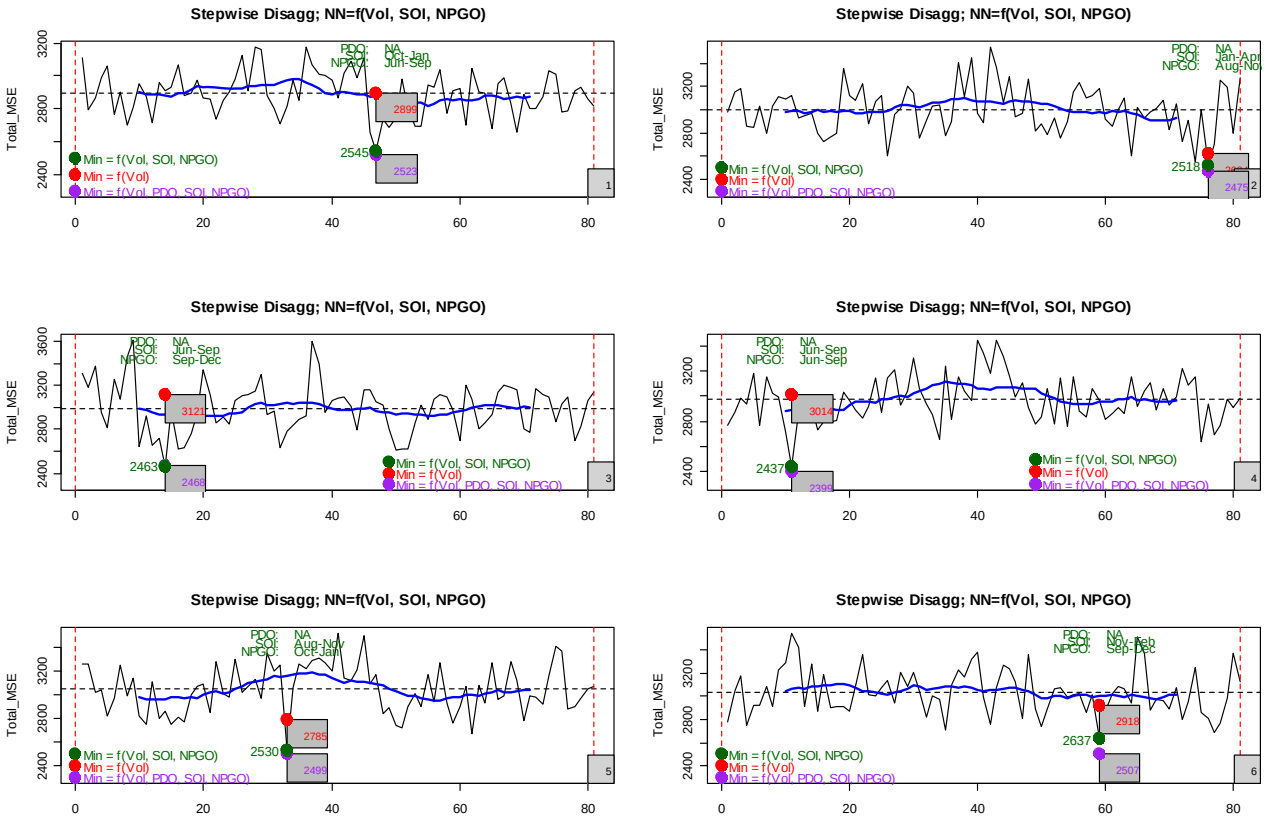

Figure 14. Same as Figure 12 but for **Y** = (*Vol, SOI, NPGO*). Number of combinations of seasonal periods = 81, instead 729. Results of minimum TMSE for **Y** = (*Vol, PDO, SOI, NPGO*) are also included (purple point).

Alternative 1, **Y** = (*Vol, PDO, SOI, NPGO*), was the combination that presented the best performance. However, **Y** = (*Vol, SOI, NPGO*) (Figure 14) can be considered a best vector in alternative 1 because it not only shows smaller values of minimum TMSE than **Y** = (*Vol, PDO, SOI*) and performs almost identically to **Y** = (*Vol, PDO, SOI, NPGO*), but more importantly uses less information than the latter by excluding the PDO. The latter suggests an apparently more valuable contribution of the NPGO to the MuNDi-S performance than from the PDO. Similar results were obtained from alternatives 3 and

4 (see Figures 15 – 17), i.e., incorporating the underlying conditional fluctuations in the model, but without a stepwise scheme, also improves the overall performance of the disaggregation approach. As expected, performance results from subsets with only climatic indices in **Y** (not shown) are poor when compared with groups that include the streamflow variable.

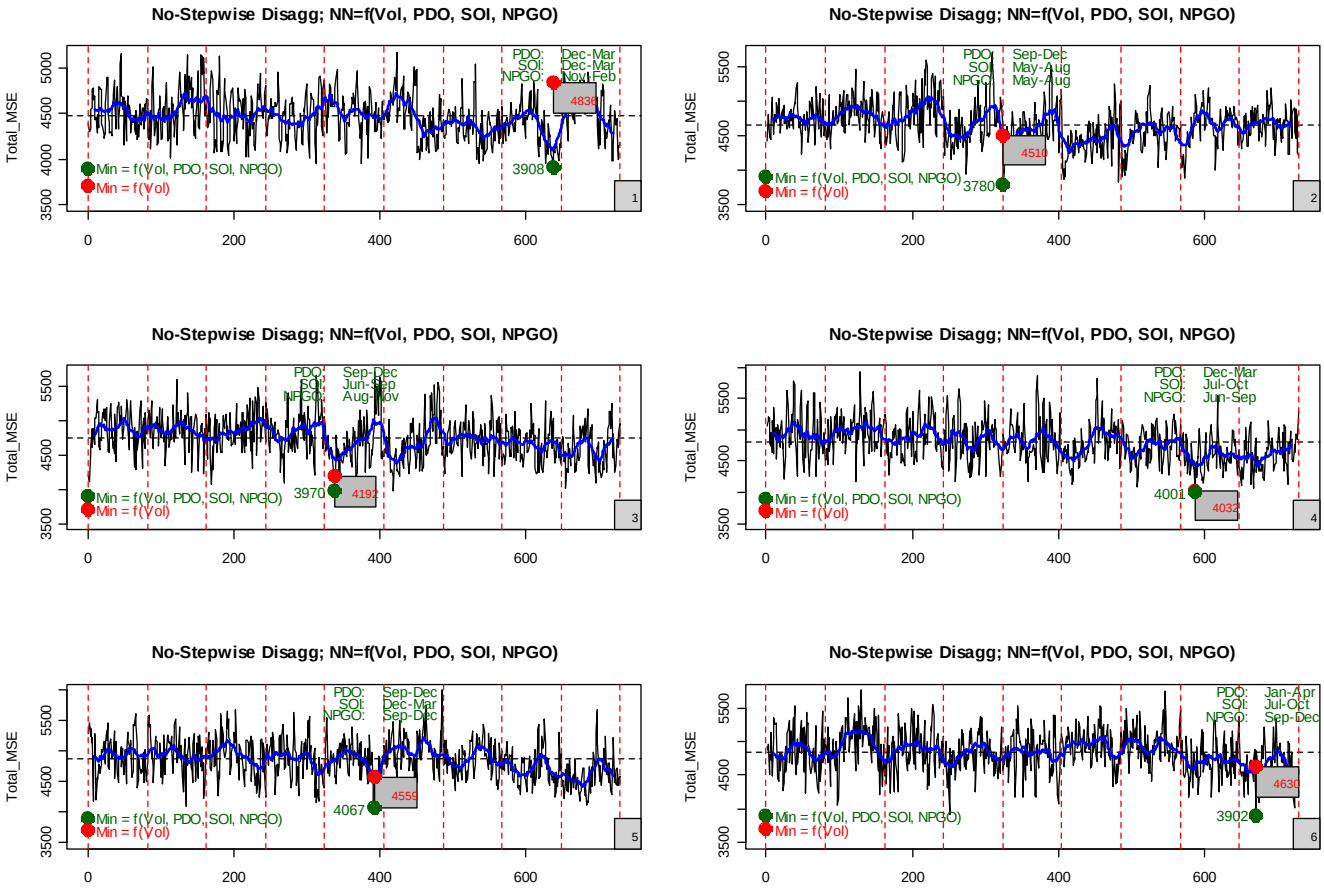

Figure 15. Same as Figure 12 but without stepwise scheme.

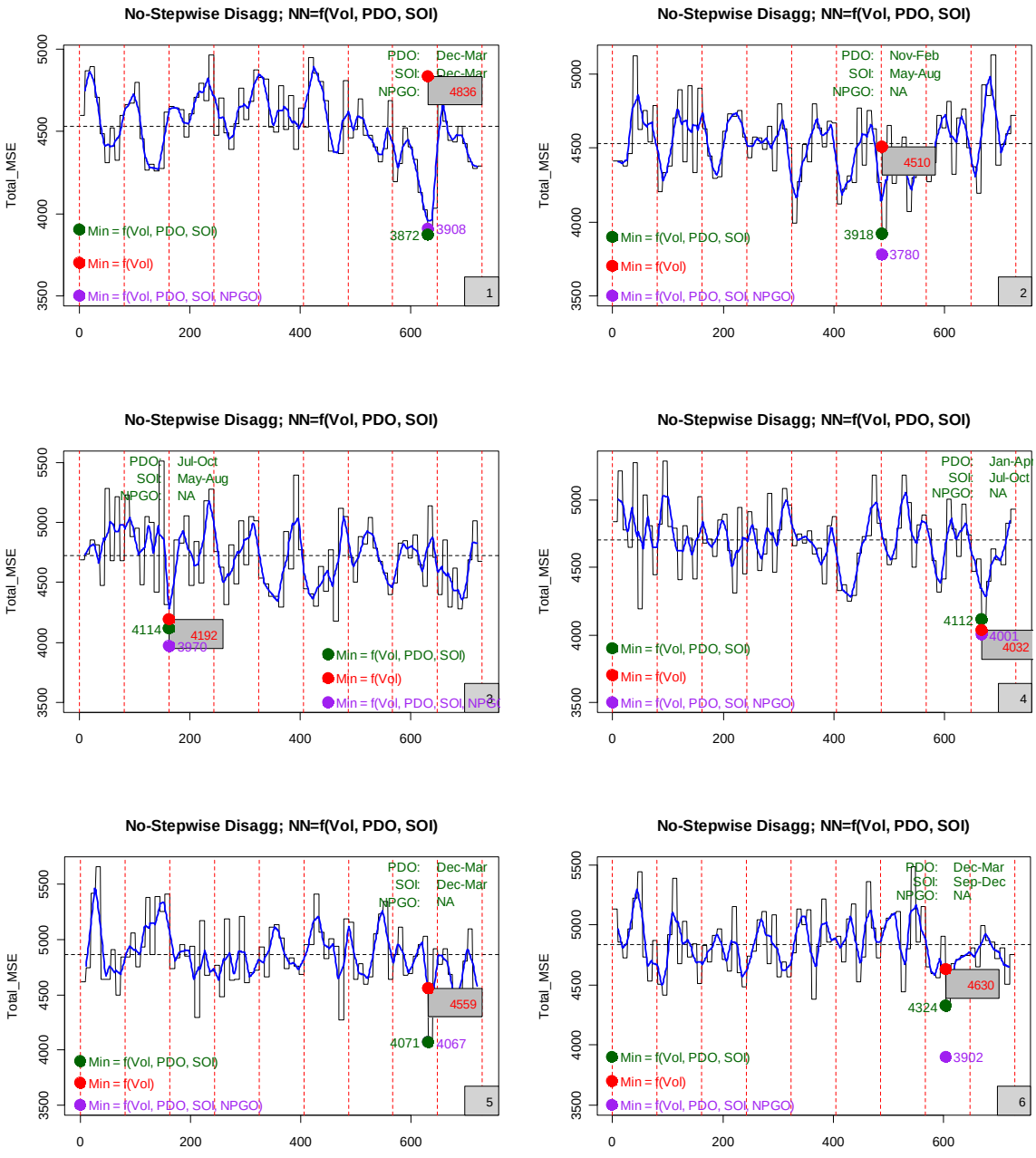

Figure 16. Same as Figure 13 but without stepwise scheme.

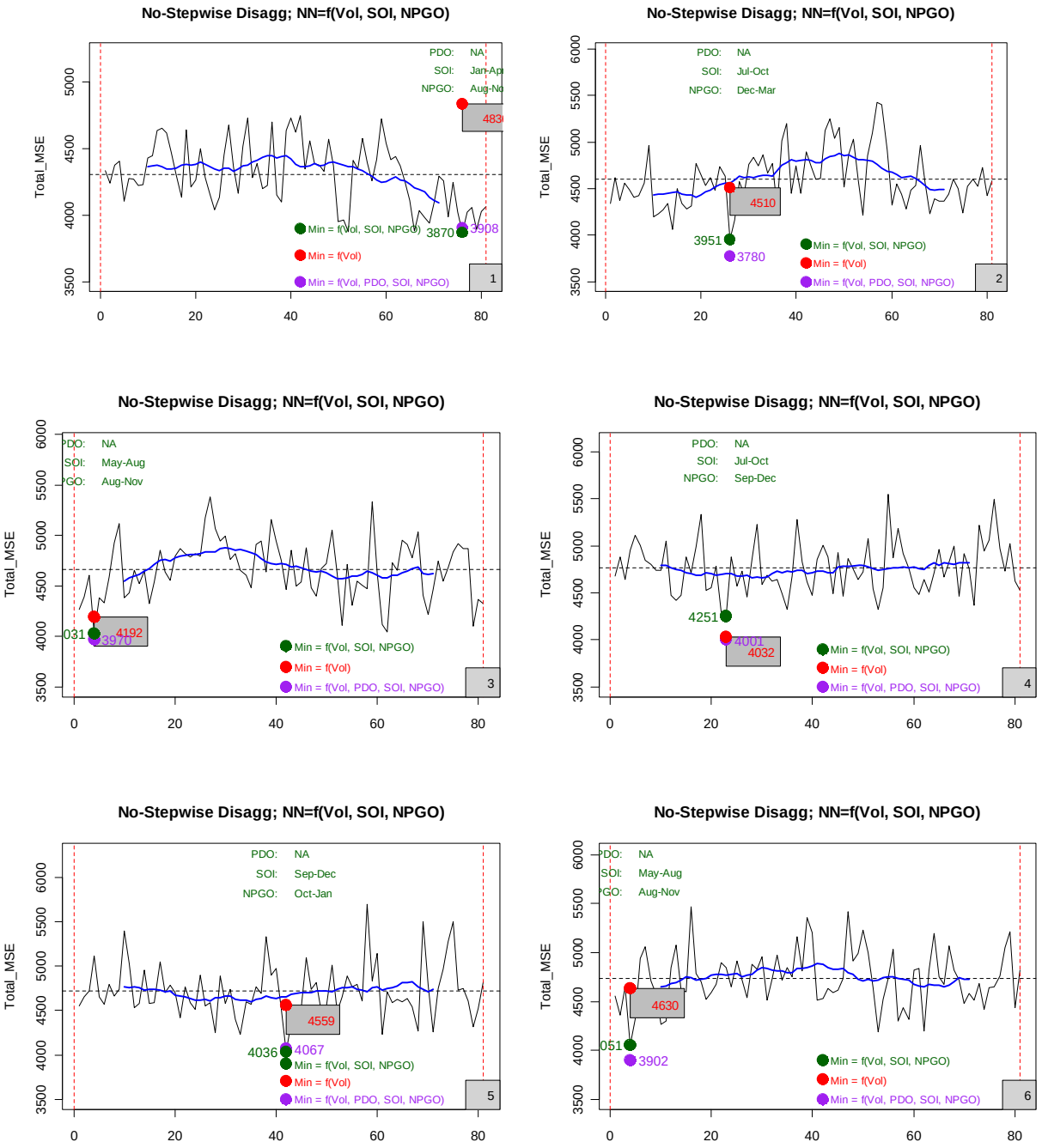

Figure 17. Same as Figure 14 but without stepwise scheme.

Considering alternative 4 as the reference model and summarizing results from the TMSE minimization, we estimated the percentage improvements in the other modeling procedures (Tables 1 - 2). Positive values mean increased performance. Our results indicate that incorporating climate signals alone to alternative 4 can produce disaggregation improvements up to 24% (Table 1).

Table 1. Percentage improvements of alternative 3 with respect to alternative 4, for spring to daily disaggregation of flows at Payette River using LOO-CV combinatorial.

Alternative	Vector ***Y*** conditional on:	k^{th} Nearest Neighbor						
		1	2	3	4	5	6	7
4	Vol							
3	Vol, PDO, SOI, NPGO	19	16	5	1	11	16	17
	Vol, PDO, SOI	20	13	2	-2	11	7	14
	Vol, PDO	23	14	-1	0	5	8	12
	Vol, SOI	15	10	4	0	4	5	12
	Vol, PDO, NPGO	24	12	0	-3	10	17	14
	Vol, SOI, NPGO	20	12	4	-5	11	13	16
	Vol, NPGO	16	13	3	-7	13	2	11

The best performance is observed for the first nearest neighbor and the worst for the fourth one. The subset **Y** = (*Vol, PDO, SOI, NPGO*) is the best combination. Most remarkable in alternative 3 is that including at least one index additional to flow in **Y** can provide comparable improvements to those when including the three signals. On the other hand, disaggregating from flows alone but including the two-level cascade approach increases significantly the model performance with respect to alternative 4 by between 25 and 40% (Table 2).

Table 2. Percentage improvements of alternatives 2 and 1 with respect to alternative 4, for spring to daily disaggregation of flows at Payette River using LOO-CV combinatorial.

Alternative	Vector ***Y*** conditional on:	k^{th} Nearest Neighbor						
		1	2	3	4	5	6	7
2	Vol	40	42	26	25	39	37	39
1	Vol, PDO, SOI, NPGO	48	45	41	40	45	46	49
	Vol, PDO, SOI	46	43	38	35	43	45	46
	Vol, PDO	40	41	33	34	37	43	47
	Vol, SOI	43	41	33	34	40	37	44
	Vol, PDO, NPGO	42	44	36	35	47	44	45
	Vol, SOI, NPGO	47	44	41	40	45	43	45
	Vol, NPGO	45	36	34	32	40	39	42

The improved performance of the stepwise disaggregation appears to be linked to the increased bootstrapping, and thus the ability of the model in decreasing the mean error of the predicted flows. The best improvement when implementing alternative 1 is between 40 and 50% approximately (Table 2), when using both the three climatic indices or the SOI and NPGO alone additional to flow. Less improvement is achieved when SOI and PDO are added to the flow vector (between 35 and 46%). From this it follows that NPGO improves the performance in the stepwise disaggregation more than the PDO when accompanied by SOI. Finally, including the PDO in **Y** = (*Vol, SOI, NPGO*) is not useful in the model performance when applied to our study region. The first issue is likely attributed to the independence found between NPGO and SOI in a seasonal time scale. These analyses support earlier arguments and suggestions presented above about

the utility of using ENSO and NPGO instead ENSO and PDO in western U.S. operational hydrology.

It is clear that alternative 1 represents the optimal procedure, followed by alternatives 2, 3, and 4. However, regarding simplicity in the model, it may be more flexible and preferable to implement alternative 2. We consider alternative 2 to have significant advantages over alternative 1, and obviously over alternative 3, because the first produces large improvements, needs less information and less steps (e.g., PCA is not required), and thus, becomes computationally less demanding. The effects of implementing the stepwise disaggregation on the preservation of historical sample statistics are discussed in the next section. If additional improvement is desired, the MuNDi-S model as evaluated in alternative 1 and including preferably at least two ocean-atmosphere signals in **Y** (SOI and NPGO are suggested) should be implemented, but at an extra computational cost. The recommended season subsets to be used for each signal and nearest neighbor in Payette are shown in Figures 13 – 17 (green legends).

4.5 Preservation of Statistical Properties

In this section we focus on evaluating our model for the alternative 2 mentioned in the previous section, and demonstrating the utility of the Lag-1-R correction algorithm when incorporated to alternative 2. The monthly simulations of MuNDi-S using flows alone are presented in Figures 18 – 19. All the statistics in all the months are very well preserved (Figure 18) although with a very slight tendency to underestimate the standard deviation and the extreme values in June. A large variability of the simulated low-flows in May is also observed. Negative values are not generated by MuNDi-S from positive

aggregate flows. On the other hand, multimodality and non-Gaussian features exhibited by historical monthly PDFs are very well reproduced (Figure 19). It can be seen, for example, how the large skew to the left in July (the largest between spring months; see skew plot in Figure 18) or the two modes of the historical flow in May (the first one around 620Mm^3/d, and the second one around 930 Mm^3/d) are very well captured by the simulations.

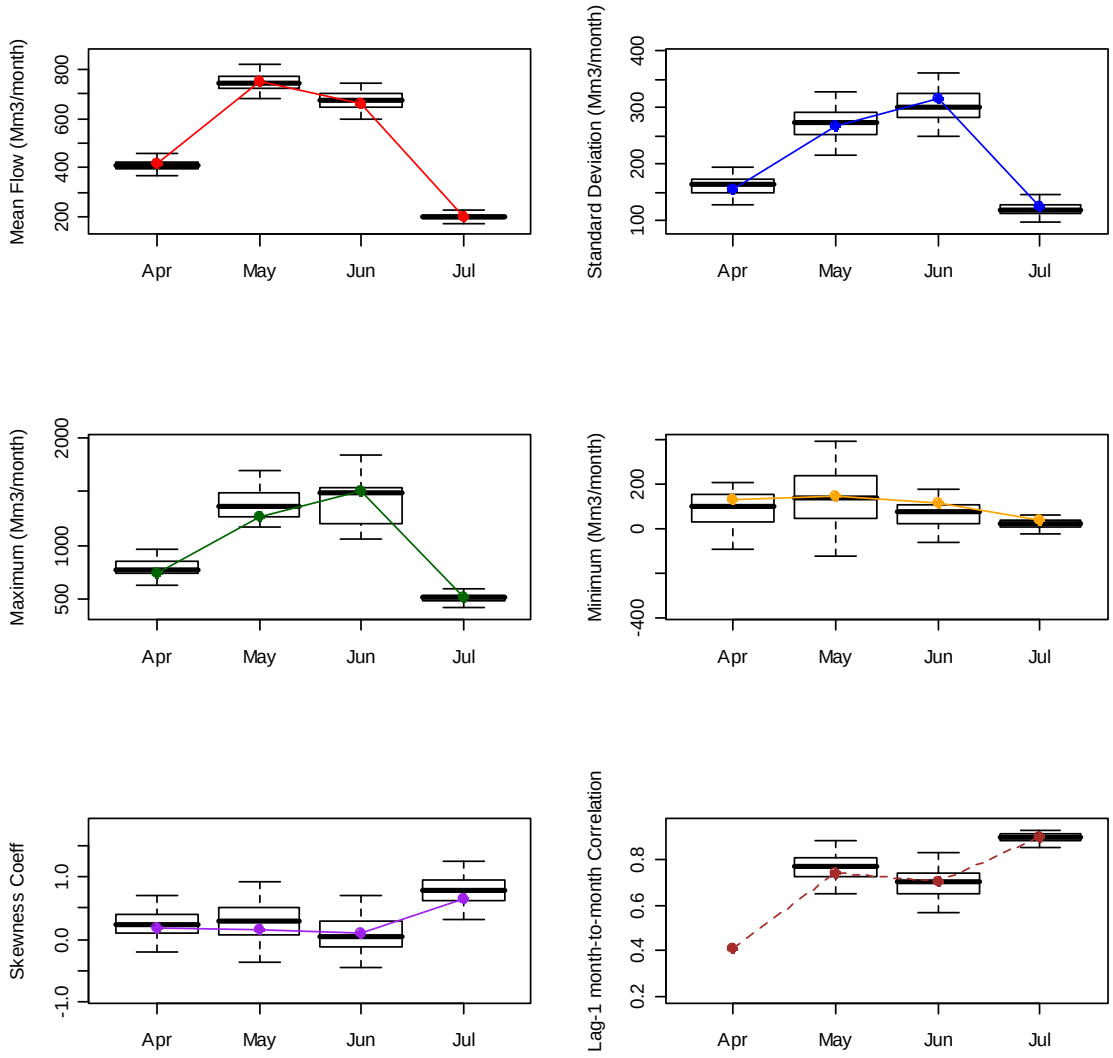

Figure 18. Preservation of historical statistics of spring flows at the monthly time scale using MuNDi-S conditional on flows alone at Payette River, Idaho. Colored Points represent the historical statistics and box-and-whisker plots the simulations.

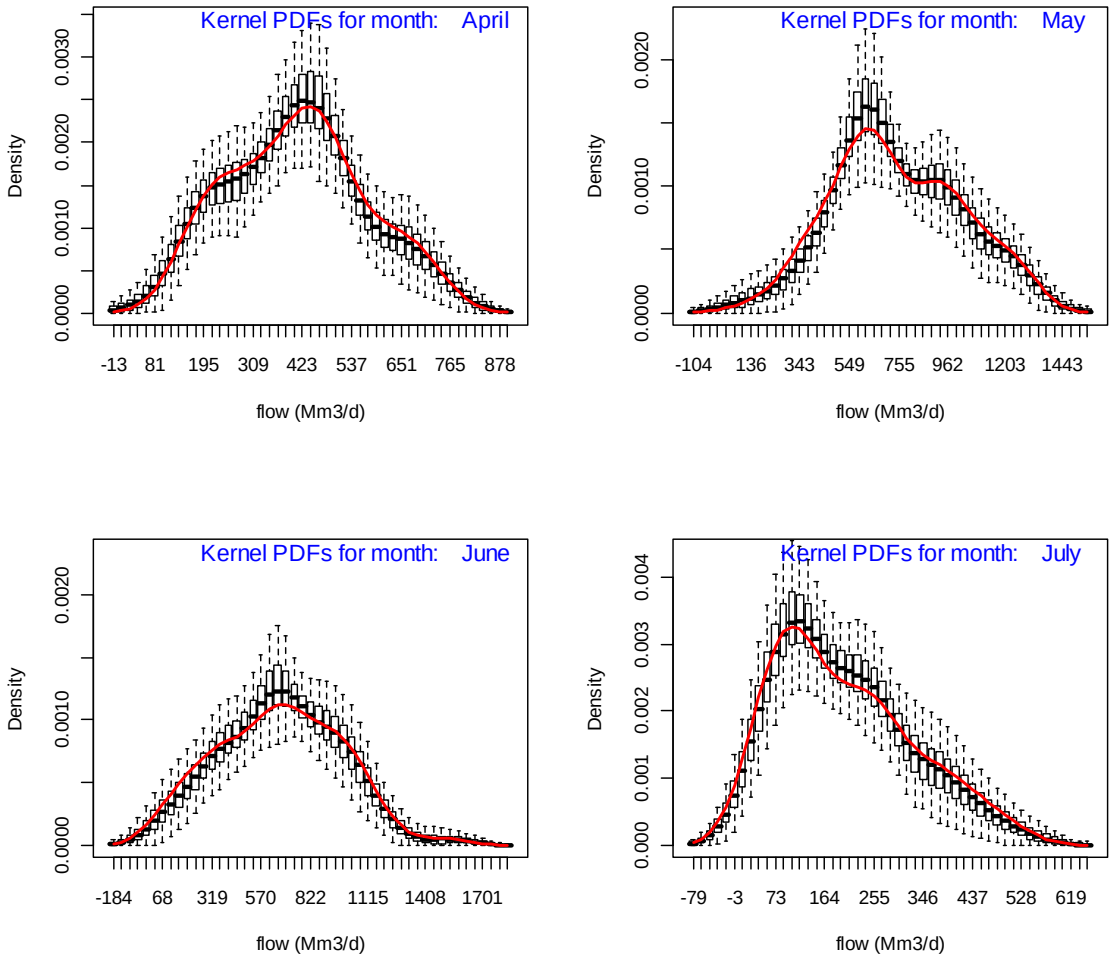

Figure 19. Preservation of historical PDFs of spring flows at the monthly time scale using MuNDi-S conditional on flows alone at Payette River, Idaho. The PDF is computed using Kernel density estimation. Red curves represent the historical PDF and box-and-whisker plots the simulations.

Regarding the second level of the disaggregation (i.e., monthly to daily), Figures 20 – 21 show results from the model evaluation. We present first the reproduction of distributional attributes and we focus specifically on May, which has the highest flows of

the year. In Figure 20 we compare the simulated and observed marginal density estimates of daily flows for the period May 10 – 21, which correspond to Julian days 130 through 141, respectively. As in the monthly simulation, historical multimodality and skewness characteristics in the month-to-days disaggregation are very well captured by the synthetic PDFs. We found that MuNDi-S reproduced all daily PDFs very well (122 in total), so for brevity we have presented only results for the mentioned days in May. The overall performance of MuNDi-S in reproducing observational daily statistics (Figure 21) is considered to be between good and very good, although some considerations and limitations are outlined below. Estimates of the standard deviation, skewness, and maxima underestimate slightly the historical for a few days in the second half of June. Low-flow daily values in May appear with a wide span in the IQR, as observed for the same month in the monthly generation. Although the lag-1 correlation is very well preserved for all days, the model is unable to simulate adequately the correlation between April 30th and May 1st, May 31st and June 1st, and in a minor degree between June 30th and July 1st (Figure 21). This limitation of the model is to the result of the stepwise scheme in the second level of the disaggregation, given that the conditional PDF in a specific month does not include dependence features from the previous month. Such specific cases represent only a small fraction of the total generated for the whole season (3/121 = 2.5%). At each level of the cascade scheme the model guarantees conservation of mass.

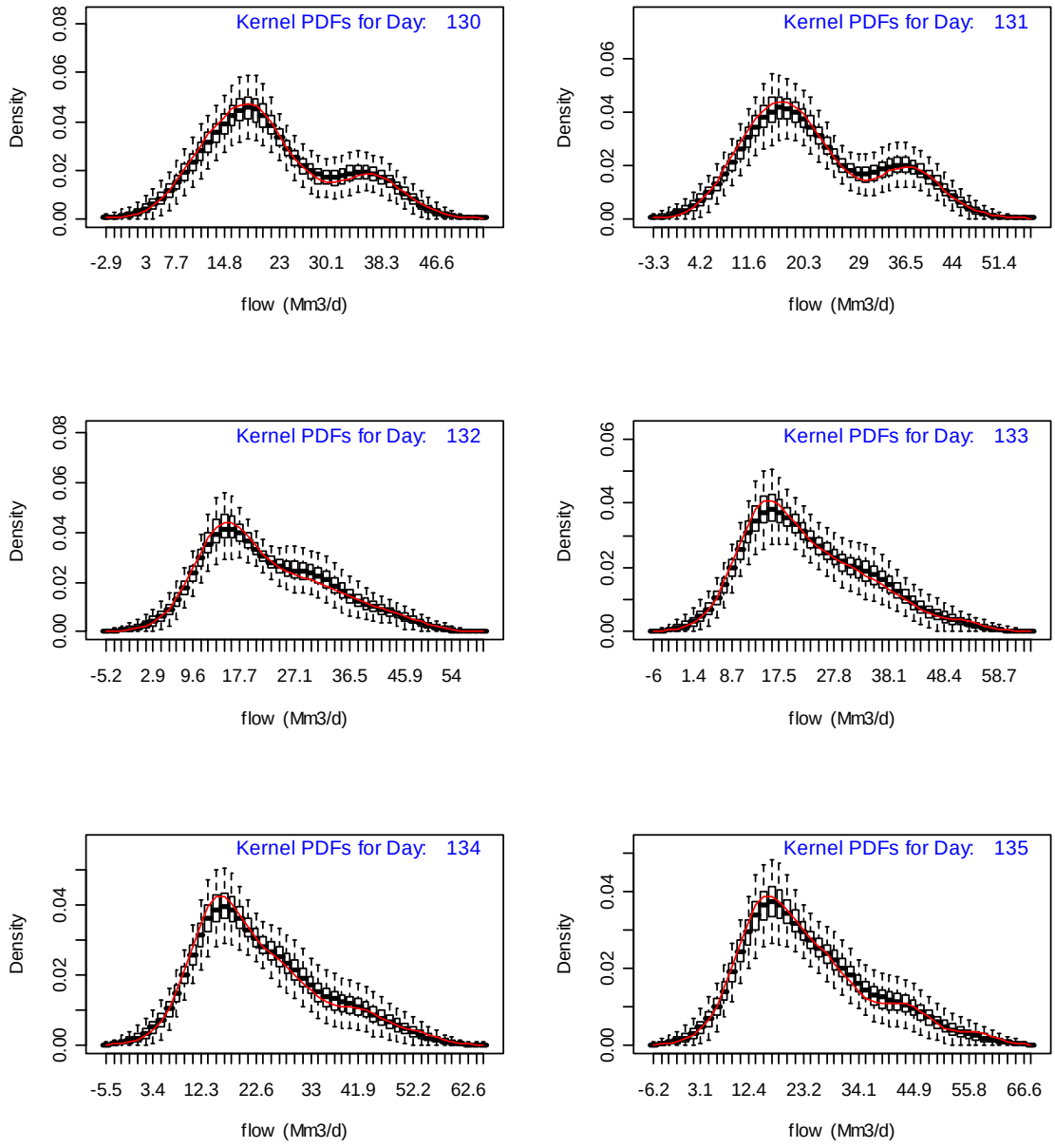

Figure 20. Same as Figure 19 but for spring flows at the daily time scale. The depicted Julian days within the plots correspond to days in the period May 10 – May 21.

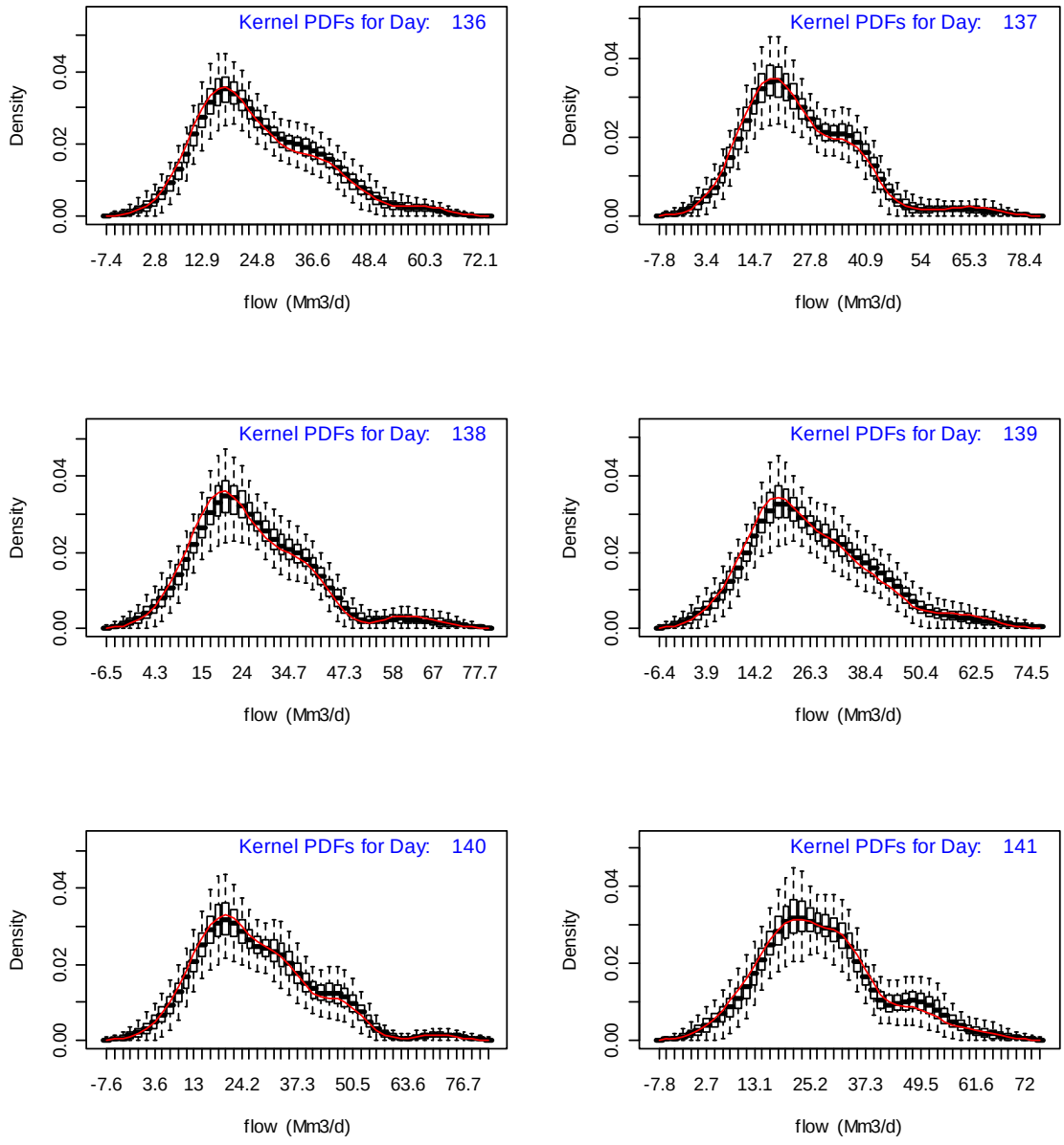

Figure 20. (Continued). Same as Figure 19 but for spring flows at the daily time scale. The depicted Julian days within the plots correspond to days in the period May 10 – May 21.

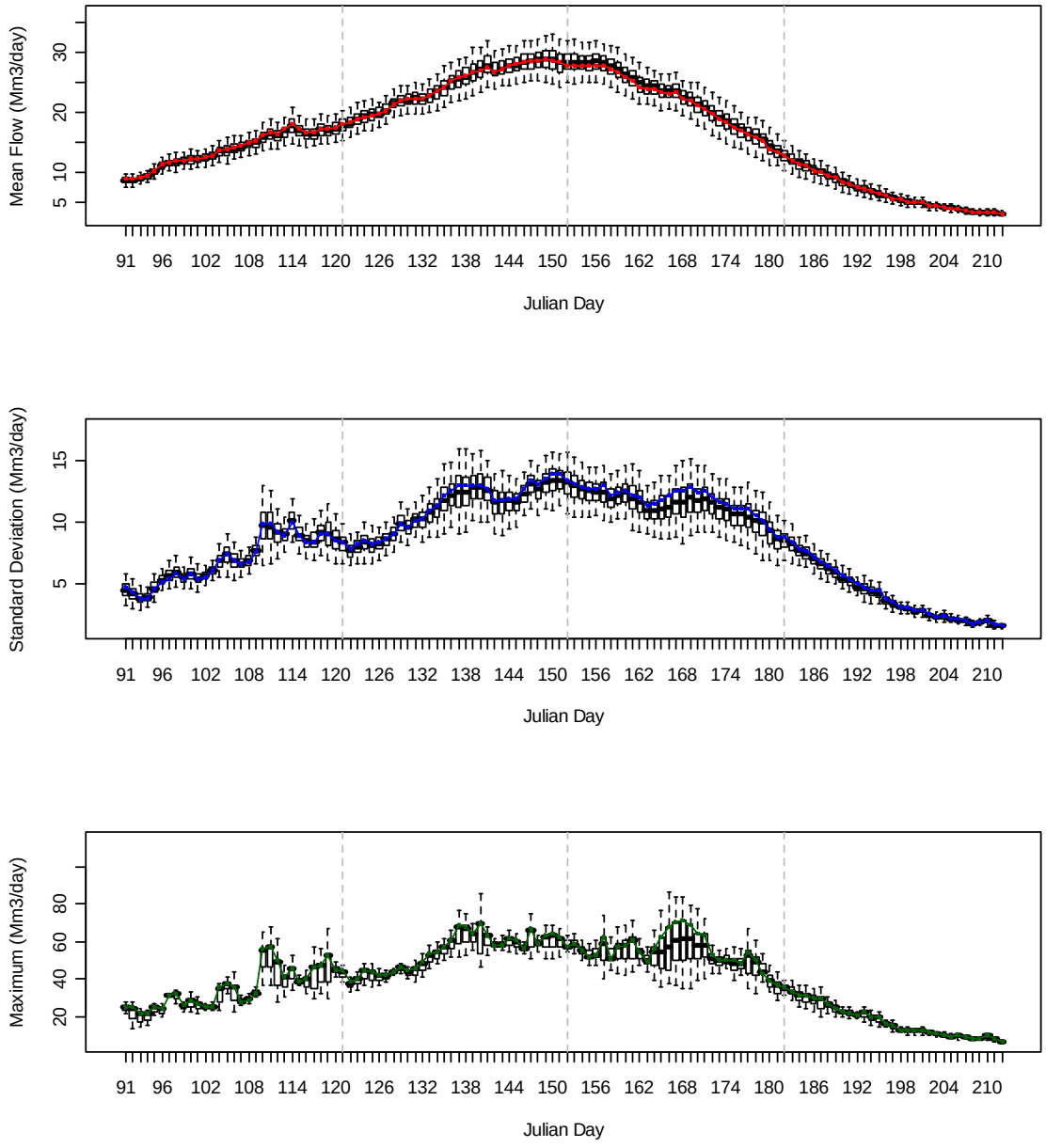

Figure 21. Same as Figure 18 but for spring flows at the daily time scale. Note that the lag-1 correlations in May 1st, June 1st, and in a minor grade July 1st (green boxes) are not preserved by MuNDi-S.

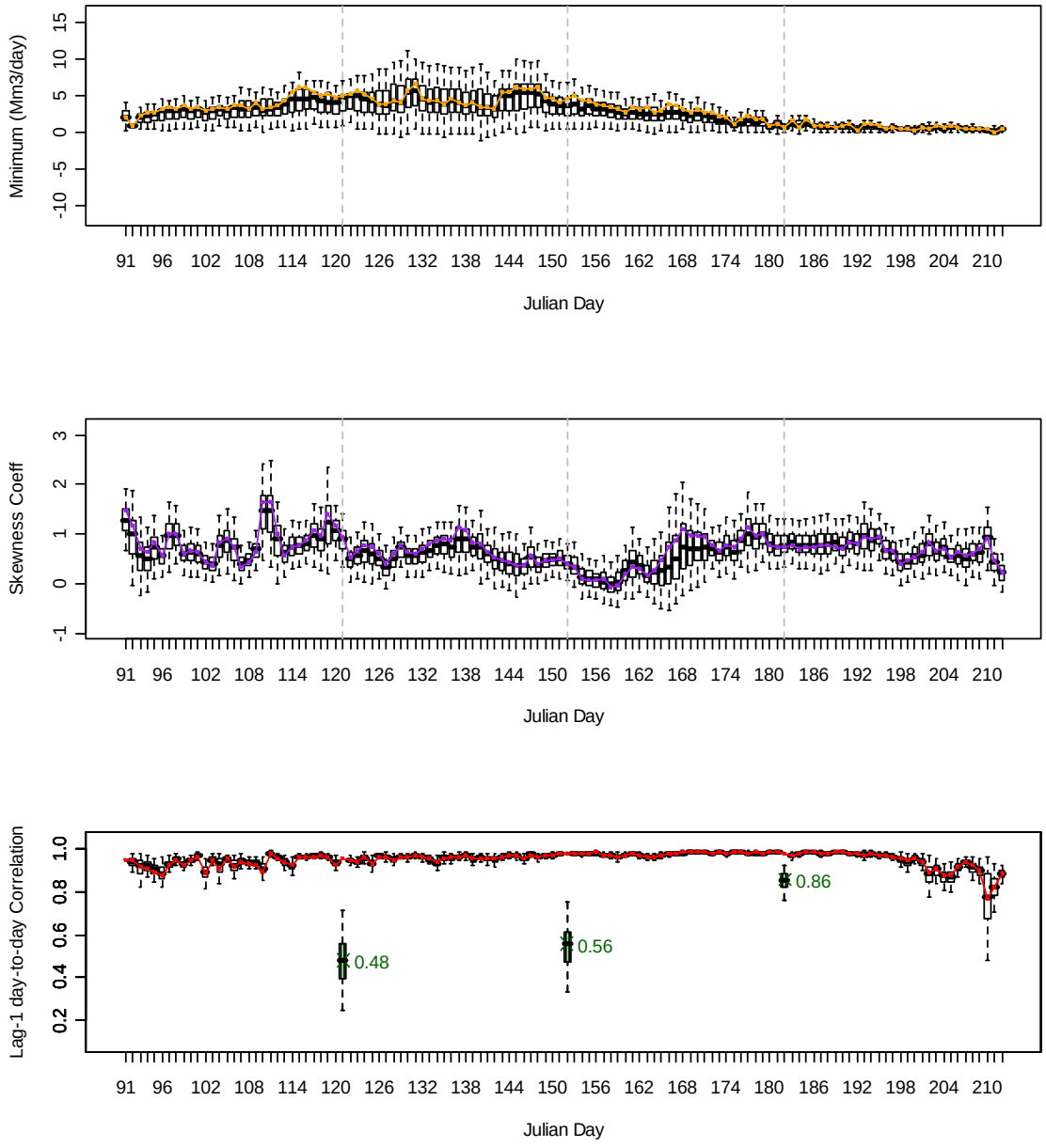

Figure 21. (Continued). Same as Figure 18 but for spring flows at the daily time scale. Note that the lag-1 correlations in May 1st, June 1st, and in a minor grade July 1st (green boxes) are not preserved by MuNDi-S.

In order to improve the preservation of the autocorrelation between consecutive days of two different months, we implemented the so called lag-1-R correction algorithm. As shown in Figure 22, MuNDi-S with lag-1-R correction improved importantly the preservation of the correlation between flows of adjacent days in different months, without affecting preservation of other daily autocorrelations.

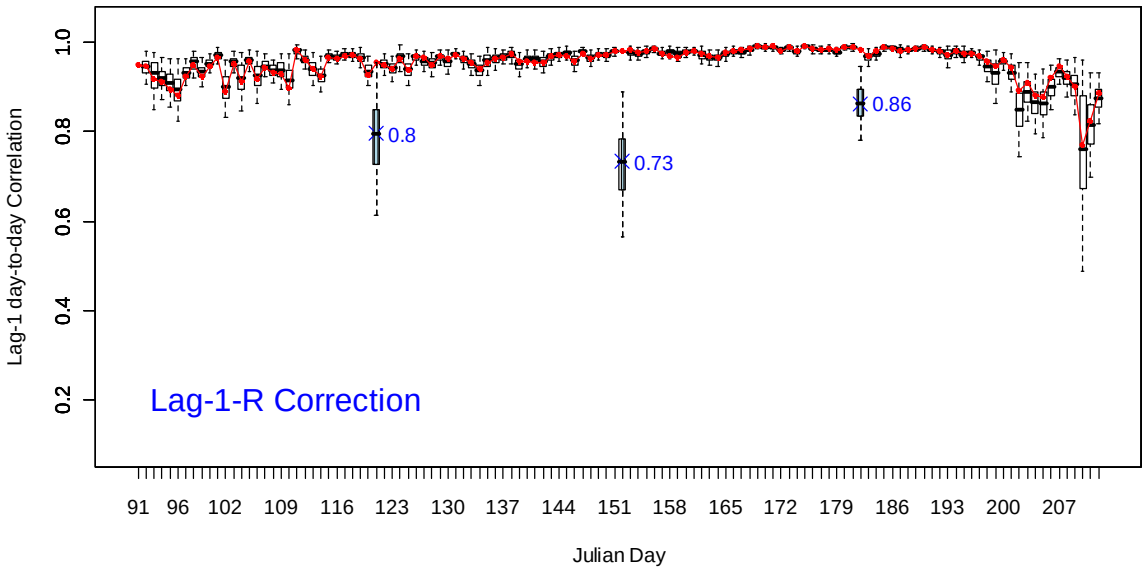

Figure 22. Preservation of historical lag-1 autocorrelation of spring flows at the daily time scale incorporating Lag-1-R correction to MuNDi-S conditional on flows alone at Payette River, Idaho. Red points represent the historical statistics and box-and-whisker plots the simulations. The improved correlations in May 1st and June 1st are depicted in blue.

The median of the simulated lag-1 *R* for May 1st was improved from 0.48 to 0.80 and for June 1st from 0.56 to 0.73. The lag-1 *R* on July 1st remained constant at 0.86. Perturbations in other historical sample statistics resulting from this procedure were minor. As can be seen from Figure 23, performance of MuNDi-S with lag-1-R correction is practically indistinguishable from that of the original model (Figure 21). Only a very slight underestimation is observed for a few days of June for the standard deviation, maximum and skewness. The lag-1-R correction is only applied at the monthly-daily cascade level, and thus, statistics and dependencies at the monthly time scale are not affected. PDFs are also very well reproduced when incorporating the mentioned nonlinear adjustment.

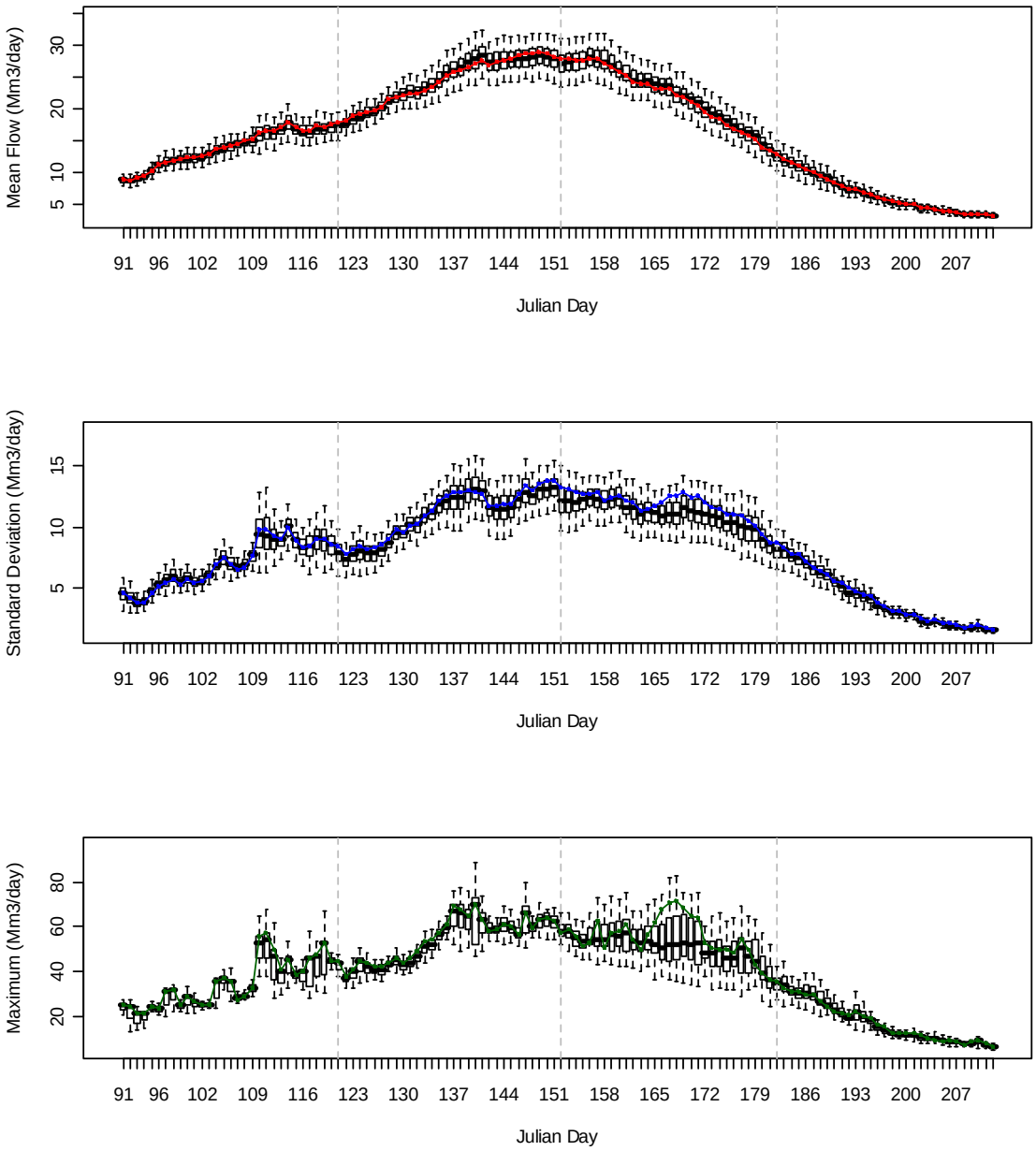

Figure 23. Preservation of historical statistics of spring flows at the daily time scale incorporating Lag-1-R correction to MuNDi-S conditional on flows alone at Payette River, Idaho. Colored points represent the historical statistics and box-and-whisker plots the simulations.

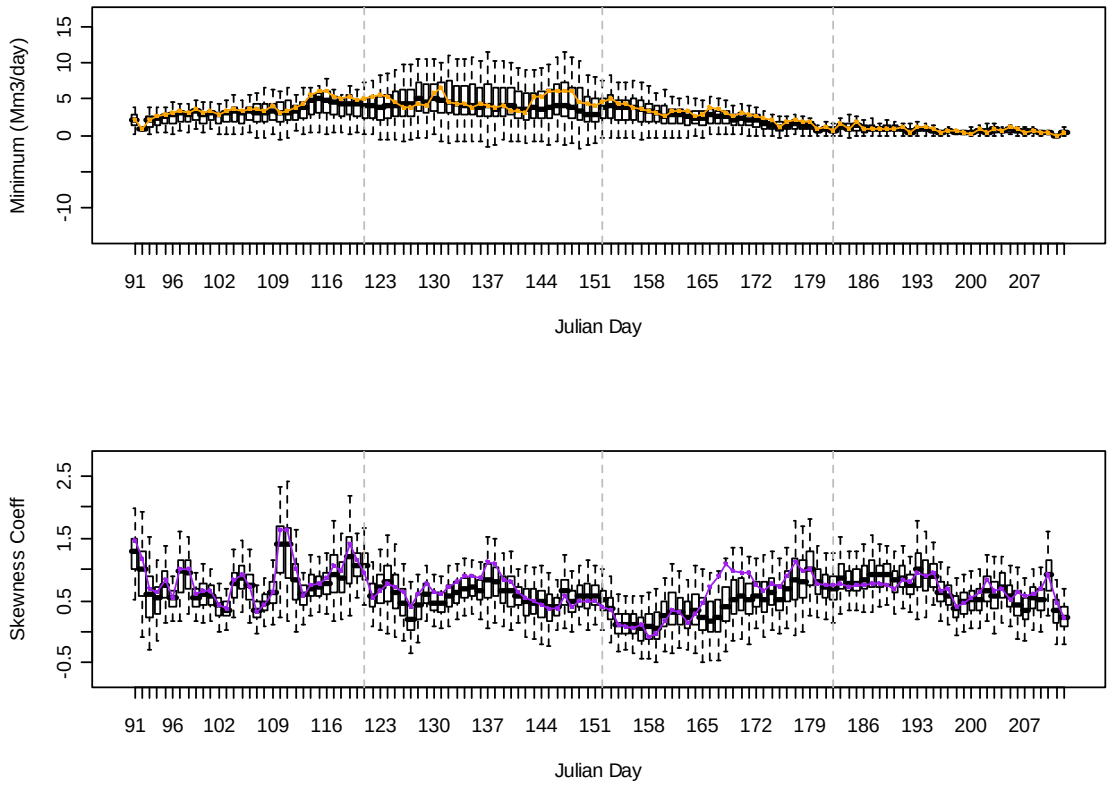

Figure 23. (Continued). Preservation of historical statistics of spring flows at the daily time scale incorporating Lag-1-R correction to MuNDi-S conditional on flows alone at Payette River, Idaho. Colored points represent the historical statistics and box-and-whisker plots the simulations.

5 Discussion and Conclusions

A multivariate nonparametric framework for stepwise disaggregation of seasonal runoff volumes to daily streamflow was presented and evaluated. The model is data driven, stochastic and purely nonparametric. This non-parametric framework allows consideration of the random nature and nonlinearities exhibited by daily flows without the need to make *a priori* assumptions about the nature of the statistical distribution of the flows. We extended relevant works on nonparametric disaggregation (Tarboton et al., 1998; Prairie et al., 2007) to: (1) exploit the potential of downscaling by stages, (2) incorporate ocean-atmosphere oscillations, (3) downscale to finer time scales, and (4) integrate PCA and KNN to bootstrap from the multivariate conditional PDF. Our contributions also include a procedure based on nonparametric resampling for improving the preservation of the autocorrelation. In addition to the improvements in the preservation of the correlation structure, our new framework does not produce negative values, guarantees conservation of mass, and preserves quite well statistical and distributional features and time dependence structures of historical data. By using a validation strategy based on a LOO-CV combinatorial procedure, we present evidence showing that it is possible to reach significant improvements in decreasing the mean error of the synthetic flow traces under alternative simulating approaches. For that, this thesis introduces and explores in detail the incorporation of climatic signals in the conditional PDF as well as the implementation of the piece-wise cascading scheme in the proposed parsimonious downscaling model as valuable tools to improve model performance.

Following, some final remarks about our work and suggestions for future research are outlined.

The improved ability of the MuNDi-S approach was demonstrated through the incorporation of large-scale climatic variables in the downscaling algorithm, confirming the fact that large-scale climatic phenomena play an important role in determining the flow patterns of the study area, even at a daily time scale.

We found significant lagged cross-correlation between climate fluctuations and spring flows in the study basin. Correlations between spring flows in the Payette River basin and SSTs during the fall and winter both in the North Pacific (between 40°N and 50°N) and the tropical Pacific (between equator and 10°N) exhibit the highest values and longest memories with respect to those cross-correlations located in other oceanic regions. SLP-based analyses in the tropics suggest that the preceding summer season in this region plays a more important role in the spring runoff variability in the study basin. In a first attempt to evaluate our model, we used indices of the PDO, ENSO, and NPGO patterns additional to flow as conditioning variables. The PDO, SOI, and NPGO signals were found to be significantly correlated with spring runoff volumes. However, it is possible to look for more specific regional fluctuations in the ocean-atmosphere system that account for more variance in the streamflow response. If included in MuNDi-S, we expect to get additional improvements in the model performance. The maps of correlations that we generated in this work provide the first step in that direction. Regarding this issue, our statistical analysis suggests that SST-based regional indices in the central Pacific between 160°W and 165°W, and in the North Pacific at the region 178°W-182°W with 40°N- 50°N (Figure 5), may be developed in future research and

explored as additional inputs for implementing MuNDi-S for the Payette River Basin. The ENSO index for the Niño 3.4 region, for example, may be a very useful input additional to the SOI. A combined SOI/Niño 3.4 index might be a best indicator and more robust measure of the ENSO dynamics (Sobolowski and Frei, 2007). Further investigation of these issues dealing with the proposed generation scheme is suggested. Other key information (e.g., predictions of local and seasonal temperatures and/or seasonal precipitation falling after a runoff volume forecast issue date, etc.) can be further included in the conditioning vector. Similarly, further analyses involving other long-lead climatic seasons and phase relationships between the studied indices should be carried out in order to define the best subset of phases with more impact both in the hydroclimatology of the study region and the model performance. Regarding the latter in particular, the SOI together with the NPGO were found to provide better improvements of performance of MuNDi-S than SOI and PDO. We attribute this to the statistically significant independence observed between fall-winter NPGO and spring-summer SOI, whereas correlations between SOI and PDO appear statistically significant in all seasons. According to Newman et al. (2003), the PDO is dependent on ENSO on all time scales. The spectral analysis described in this thesis identified coherences between the evaluated signals and verified that the NPGO exhibits more prominent low-frequency oscillations than the PDO. We identified a connection between the NPGO and the SOI in the spectral domain, where the decadal variability of the SOI signal appears modulated by that of the NPGO.

In the case of the stepwise scheme, where MuNDi-S was considerably better than models without cascades, its role in incorporating more neighbors in the disaggregation

process and the increased bootstrapping appear as the main explanation of the model performance improvement. A two-level symmetric cascade scheme was considered in the model, in which a spring runoff volume is disaggregated into monthly values first, and then, each monthly volume is disaggregated into daily components. The downscaling algorithm implemented here has the ability to produce daily flows that have not been seen in the past while generating sequences that are statistically indistinguishable from the observations. Our selection of an appropriate model for the aggregate space (i.e., spring volumes with AR-1) plays also an important role in achieving this goal given the capabilities inherent to parametric approaches.

One limitation of our model resulting from the stepwise procedure is the lack of preservation of historical correlation between daily flows across months. This limitation is common also to parametric downscaling models. However, the lag-1-R correction algorithm presented in this thesis was successfully incorporated and tested in MuNDi-S and the correlation on those days was importantly improved. As is normal in this type of models, one concern with MuNDi-S is related with the amount of data/processes required, depending on, (1) the higher dimensionality both in the conditioning information and the number of observations, and (2) the increased bootstrapping both in the stepwise scheme and the Lag-1-R correction algorithm, if implemented. These factors cause growing computational needs, and even more when extended to space-time disaggregation, but this may not be a problem considering the current processing and storage capabilities of even personal computers. In the model evaluation presented here regarding reproduction of historical statistics, only flows were considered in the conditional vector. Synthetic generation of annual, seasonal and monthly series of PDO,

SOI and NPGO indices (or additional variables) that preserve statistics, dependencies and distributional features in the multivariate aggregate space was beyond the scope of this work.

Finally, we suggest exploring the application of the streamflow disaggregation methods presented in this thesis to other regions with different hydroclimatic conditions (e.g. streams with intermittent behavior) and other space and time scales. Similarly, the proposed approach should be applied and extended to stochastic modeling of other hydroclimatological processes, especially those with complex space-time dependence structures and that exhibit teleconnections with ocean-atmospheric variations.

6 Acknowledgements

Partial funding for this research by the US Bureau of Reclamation Science and Technology Program under Project ID2601 is gratefully acknowledged.

7 References

Barton, S. B., and J. A. Ramirez (2004), Effects of El Niño Southern Oscillation and Pacific Interdecadal Oscillation on water supply in the Columbia River basin, *J. Water Resour. Plan. Manage.*, 130 (4), 281-289.

Bracken, C., B. Rajagopalan, and J. Prairie (2010), A multisite seasonal ensemble streamflow forecasting technique, *Water Resour. Res*., 46, W03532, doi:10.1029.

Bras, R. L., and I. Rodriguez-Iturbe (1985), *Random Functions in Hydrology*. Addison-Wesley, Reading, Massachusetts, 559 pp.

Chebaane, M., J.D. Salas and D.C. Boes (1995), Product Periodic Autoregressive Processes for Modeling Intermittent Monthly Streamflows, *Water Resour. Res*., 31(6):1513-1518.

Chhak, K., E. Di Lorenzo, N. Schneider, and P. F. Cummins (2009), Forcing of low-frequency ocean variability in the northeast Pacific, *J. Clim*., 22(5), 1255– 1276.

Clark, M. P., M. C. Serreze, and G. J. McCabe (2001), Historical effects of El Nino and La Nina events on the seasonal evolution of the montane snowpack in the Columbia and Colorado River Basins, *Water Resour. Res*., 37, 741– 757.

Cleveland, W. S., E. Grosse, and W. M. Shyu (1992), Local regression models, chapter 8 in: Chambers, J. M. and Hastie, T. J. (eds), *Statistical Models in S*. Wadsworth and Brooks/Cole, Pacific Grove, CA, pp. 309-376.

Davidson, A. C. and D. V. Hinkley (1997), *Bootstrap Methods and their Application*. Cambridge University Press, New York, 582 pp.

Deser, C., A. S. Phillips, and J. W. Hurrell, (2004), Pacific Interdecadal climate variability: Linkages between the tropics and North Pacific during boreal winter since 1900. *J. Climate*, 17, 3109–3124.

Di Lorenzo, E., A. J. Miller, N. Schneider, and J. C. McWilliams (2005), The warming of the California Current System: Dynamics and ecosystem implications, *J. Phys. Oceanogr*., 35(3), 336–362.

Di Lorenzo, E., et al. (2008), North Pacific Gyre Oscillation links ocean climate and ecosystem change, *Geophys. Res. Lett*., 35, L08607, doi:10.1029/2007GL032838.

Di Lorenzo, E., N. Schneider, K. M. Cobb, J. C. Furtado, and M. A. Alexander (2009), ENSO and the North Pacific Gyre Oscillation: an integrated view of Pacific decadal dynamics, *Submmited to Geophys. Res. Lett*.

Efron, B., and R. J. Tibshirani (1993), *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York, 436 pp.

Fernádez, B. and J.D. Salas (1986), Periodic gamma autoregressive processes for operational hydrology. *Water Resour. Res*., 22 (10), 1385–1396.

Gershunov, A. (1998), ENSO influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: Implications for long-range predictability, *J. Clim*., 11, 3192–3203.

Gershunov, A., and T.P Barnett (1998), Interdecadal modulation of ENSO teleconnections. *Bull. Am. Meteorol. Soc*., 79(12), 2715–2725.

Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona (2005), A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts, *Water Resour. Res*., 41, W10410, doi:10.1029.

Grygier, J.C., and J.R. Stedinger (1988), Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res.*, 24(10), 1574–1584.

Hamlet, A.F., and D.P. Lettenmaier (1999), Columbia River streamflow forecasting based on ENSO and PDO climate signals, *J. Water Resour. Plan. Manage.*, 125(6), 333–341.

Hamlet, A.F., D. Huppert, and D.P. Lettenmaier (2002), Economic value of long-lead streamflow forecasts for Columbia River hydropower, *J. Water Resour. Plan. Manage.*, 128(2), 91-101.

Harms, A. A., and T. H. Campbell (1967), An extension to the Thomas-Fiering model for the sequential generation of streamflow, *Water Resour. Res.*, 3(3), 653–661.

Harshburger, B., Y. Hengchun, and J. Dzialoski (2002), Observational evidence of the influence of Pacific SSTs on winter precipitation and spring stream discharge in Idaho, *J. Hydrol*., 264(1– 4), 157– 169.

Hidalgo, H. G., and J. A. Dracup (2003), ENSO and PDO effects on hydroclimatic variation of the Upper Colorado River Basin, *J. Hydrometeorol.*, 4, 5 – 23.

Higgins, J.J. (2004), *Introduction to Modern Nonparametric Statistics*. Brooks/Cole, Pacific Grove, CA, 366 pp.

Hipel, K.W., and A.I. McLeod (1994), *Time Series Modeling of Water Resources and Environmental Systems*. Elsevier, Amsterdam, 1013 pp.

Jackson, D.A. (1993), Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches, *Ecology*, 74(8), 2204–2214.

Johnson, D.E. (1998), *Applied Multivariate Methods for Data Analysts*. Brooks/Cole Pub., Pacific Grove, CA, 567 pp.

Koch, R. W., and A. R. Fisher (2000). Effects of inter-annual and decadal-scale climate variability on winter and spring streamflow in western Oregon and Washington. *Proceedings of the 68th annual Western Snow Conference*, Port Angeles, Washington, 1-11.

Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resour. Res*., 32(7), 2105– 2117.

Kumar, D.N., U. Lall, and M.R. Peterson (2000), Multisite disaggregation of monthly to daily streamflow, *Water Resour. Res*., 36(7), 1823-1833.

Lall, U. and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res*., 32(3), 679-693.

Lane, W.L. (1979), *Applied Stochastic Techniques, Users Manual*, Eng. and Res. Cent., Bureau of Reclamation, Denver, CO.

Lee, T., J.D. Salas and J. Prairie (2010), An enhanced nonparametric streamflow disaggregation model with genetic algorithm, *Water Resour. Res*. (Accepted for publication).

Linkin, M. E., and S. Nigam (2008), The north pacific oscillation-west Pacific teleconnection pattern: Mature-phase structure and winter impacts, *J. Clim*., 21(9), 1979-1997.

Loader, C. (1999), *Statistics and Computing: Local Regression and Likelihood*. Springer, New York, 290 pp.

Mantua, N. J., S. R. Hare, J. M. Wallace, and R. C. Francis (1997), A Pacific interdecadal climate oscillation with impacts on salmon production, *Bull. Am. Meteorol. Soc*., (78), 1069– 1079.

McCabe, G. J., and M. D. Dettinger (1999), Decadal variations in the strength of ENSO teleconnections with precipitation in the western United States. *Int. J. Climatol*., (19), 1399–1410.

Mejia, J.M. and J. Rousselle (1976), Disaggregation Models in Hydrology Revisited, *Water Resour. Res*., 12(2), 185-186.

Newman, M., G. P. Compo, and M. A. Alexander (2003), ENSO-forced variability of the Pacific Decadal Oscillation, *J. Clim*., (16), 3853 – 3857.

Prairie, J.R., B. Rajagopalan, U. Lall and T.J. Fulp (2007), A stochastic nonparametric technique for space-time disaggregation of streamflows, *Water Resour. Res*., 43, W03432, doi:10.1029.

Rajagopalan, B., and U. Lall (1999), A K-nearest neighbor simulator for daily precipitation and other variables, *Water Resour. Res.*, 35(10), 3089– 3101.

Rajagopalan, B., J.D. Salas, and U. Lall (2010, in press), Stochastic methods for modeling precipitation and streamflow. In: *Advances in Data-based Approaches for Hydrologic Modeling and Forecasting.* B. Sivakumar and R. Berndtsson, Eds., World Scientific, Singapore, 17-52.

Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagona (2006), A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin, *Water Resour. Res*., 42, W09404, doi:10.1029.

Salas, J.D., J.R. Delleur, V. Yevjevich, and W.L. Lane (1980), *Applied Modeling of Hydrologic Time Series,* Water Resources Publications, Littleton, CO, 484 pp.

Salas, J.D. (1993), Analysis and modeling of hydrologic time series. In: *Handbook of Hydrology*. Edited by D. R. Maidment, McGraw-Hill, New York, 19.1-19.72.

Salas, J.D. and T. Lee (2010), Nonparametric simulation of single site seasonal streamflows, *J. Hydrologic Eng*., 15(4): 284-296

Salathé Jr, E.P., P.W. Mote, and M.W. Wiley (2007), Review of scenario selection and downscaling methods for the assessment of climate change impacts on hydrology in the United States Pacific Northwest, *Int. J. Climatology*, 27, 1611-1621.

Santos, E.G. and J.D. Salas (1992), Stepwise disaggregation scheme for synthetic hydrology, *J. Hydraulic Eng.*, 118 (5), 765–784.

Schneider, N., E. Di Lorenzo, and P. Niiler (2005), Salinity variations in the Southern California Current, *J. Phys. Oceanogr*., 35(8), 1421–1436.

Schneider, N., and B. D. Cornuelle (2005), The forcing of the Pacific decadal oscillation. *J. Climate*, 18, 4355–4373.

Shaman, J., M. Stieglitz, S. Zebiak and M. A. Cane (2003), A local forecast of land surface wetness conditions derived from seasonal climate predictions, *J. Hydrometeorol.*, 4(3), 611-626.

Sharma, A., and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resour. Res*., 38(7), 1100, doi:10.1029/2001WR000953.

Shumway, R. H., and D. S. Stoffer (2006), *Time Series Analysis and Its Applications*. Second Ed., Springer-Verlag, New York, 575 pp.

Simonoff, J.S. (1996), *Smoothing Methods in Statistics*. Springer, New York, 338 pp.

Sobolowski, S., and A. Frei (2007), Lagged relationships between North American snow mass and atmospheric teleconnection indices. *Int. J. Climatology*, 27, 221-231.

Srinivas, V. V., and K. Srinivasan (2005), Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows, *J. Hydrol*., 302, 307– 330.

Stedinger, J. R., and R. M. Vogel (1984), Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res*., 20(1), 47–56.

Stedinger, J. R., D. Pei, and T. A. Cohn (1985), A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res*., 21(5), 665–675.

Takezawa, K. (2006), *Introduction to Nonparametric Regression*. Wiley, New Jersey, 538 pp.

Tarboton, D.G., A. Sharma, and U. Lall (1998), Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res*., 34(1), 107-119.

Trenberth, K. E., and J. W. Hurrell (1994), Decadal atmosphere–ocean variations in the Pacific. *Climate Dyn*., 9, 303–319.

Trenberth, K.E. (1997), The definition of El Niño. *Bull. Am. Meteorol. Soc*. 78 (12), 2771–2777.

Valencia, D.R., and J.C. Schaake (1973), Disaggregation processes in stochastic hydrology, *Water Resour. Res*., 9(3), 580-585.

Vimont, D. J., et al. (2009), Midlatitude excitation of tropical variability in the Pacific: The role of thermodynamic coupling and seasonality, *J. Clim*, 22(3), 518-534.

Wang, C., S.-P. Xie, and J. A. Carton (2004), *Earth's Climate: The Ocean-Atmosphere Interaction*. C. Wang, S.-P. Xie, and J. A. Carton, Eds., AGU Geophysical Monograph Series 147, 414 pp.

Wang, D., Y. Liu, and D. Gu (2004), Gyre-connected variations inferred from the circulation indices in the Northern Pacific Ocean. In: *Earth's Climate: The Ocean-Atmosphere Interaction.* C. Wang, S.-P. Xie, and J. A. Carton, Eds., AGU Geophysical Monograph Series, 147:319-328.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier (2002), Long-range experimental hydrologic forecasting for the eastern United States, *J. Geophys. Res*., 107(D20), 4429, doi:10.1029.

Wood, A. W., L. R. Leung, V. Sridhar, and D. P. Lettenmaier (2004), Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs, *Clim. Change*, 62(1– 3), 189– 216.