International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, California, USA
Daniel P. Ames, Nigel W. T. Quinn, Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Quantifying ENSO impacts at the basin scale using the Iterative Input variable Selection algorithm

Ludovica Beltrame [a], Daniele Carbonin [a], Stefano Galelli [b], Andrea Castelletti [a] and Matteo Giuliani[a]

[a] *Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza L. da Vinci, 32, I-20133 Milano, Italy (ludovica.beltrame@mail.polimi.it, daniele.carbonin@mail.polimi.it, andrea.castelletti@polimi.it, matteo.giuliani@polimi.it)*

[b] *Pillar of Engineering Systems and Design, Singapore University of Technology and Design, 20 Dover Drive, Singapore 138682 (stefano_galelli@sutd.edu.sg)*

Abstract: Medium-to-long range streamflow predictions provide a key assistance in anticipating hydro-climatic adverse events and prompting effective adaptation measures. In this context, recent modelling efforts have been dedicated to seasonal and inter-annual predictions based on the teleconnection between at-site hydrological processes and large-scale, low-frequency climate fluctuations, such as El Niño Southern Oscillation (ENSO). This work proposes a novel procedure for first detecting the impact of ENSO on hydro-meteorological processes at the basin scale, and then quantitatively assessing the potential of ENSO indexes for building medium-to-long range streamflow prediction models. Core of this procedure is the adoption of the Iterative Input variable Selection (IIS) algorithm, which is employed to find the most relevant determinants of streamflow variability and derive predictive models based on the selected inputs. The procedure is tested on two different case studies, the Columbia River (US) and the Williams River (Australia), whose sensitivity to ENSO fluctuations has been documented in previous studies. Results show that IIS outcomes for both case studies are consistent with the results of previous analyses conducted with state-of-the-art detection methods, and that ENSO indexes can effectively be used in both regions to enhance the accuracy of streamflow prediction models.

Keywords: ENSO; Long-range Streamflow Prediction; Input Variable Selection; Data-driven Models; Hydrological Modelling

1 INTRODUCTION

El Niño Southern Oscillation (ENSO) is a large-scale, coupled ocean-atmosphere phenomenon occurring in the tropical Pacific Ocean, and is considered one of the most significant factors causing hydro-climatic anomalies throughout the world [Kahya and Dracup, 1993]. Since the time-lag between ENSO fluctuations and at-site hydrological processes is typically of the order of few months, the accuracy and lead-time of long-range streamflow prediction models is potentially enhanced by including ENSO indexes among the predictors.

Different methods to investigate the teleconnection between ENSO and at-site hydrological processes have been proposed during the past two decades. The harmonic analysis, originally described by Ropelewski and Halpert [1986], is one of the most adopted. For example, Chiew and McMahon [2002] use it for detecting teleconnection and quantifying its strength across many geographical regions. Other commonly adopted methods are correlation analysis (see, among the others, Gutierrez and Dracup [2001]; Hidalgo and Dracup [2003]; Opitz-Stapleton et al. [2007]) and cross-lagged correlation [Chiew et al., 1998]. Statistical tests are also used to assess the significance of the ENSO-streamflow rela-

tionships. For example, Simpson et al. [1993] and Kiem and Franks [2001] use a chi-square test and a Student's *t*-test on two river basins in South-East Australia affected by ENSO fluctuations. In spite of their widespread use, these methods have two main shortcomings: *i*) correlation and cross-lagged correlation analysis are based on the restrictive assumption of a linear relationship between ENSO and streamflow variability, whereas hydro-climatic processes are highly non-linear [Sharma, 2000]; *ii*) the positive contribution of ENSO indexes (e.g. Multivariate ENSO Index, Niño 3.4 SST, Southern Oscillation Index; see Kiem and Franks [2001] and references therein) to the predictive accuracy of streamflow prediction models is detected but not quantified. In fact, developing a quantitative assessment of ENSO effects on a certain area would mean not only understanding whether ENSO indexes could usefully be employed for streamflow forecasting purposes, but also exactly determining which ENSO indexes and corresponding time-lags would maximise the accuracy and lead-time of a predictive model.

This paper contributes a novel procedure based on Input Variable Selection (IVS, e.g., Guyon and Elisseeff [2003]) to detect and quantify ENSO-hydro-climatic teleconnection for predictive purposes. The procedure is composed by the following four steps: i) identification of the ENSO signal for the specific river basin under investigation; ii) characterisation of the hydrological response to ENSO; iii) assessment of statistical significance; iv) application of IVS to assess the potential of ENSO indexes for building medium-to-long range streamflow prediction models. IVS is employed to provide useful quantitative information about the forecast potential of both meteorological and ENSO indexes, thus identifying the best inputs to a long-range streamflow prediction model and the associated time-lags. The IVS-based procedure is demonstrated on two case studies, the Columbia River basin, located in the western US, and the Williams River basin, in eastern Australia. The two basins are characterised by different climate conditions and spatial scales: this allows verifying whether the proposed procedure is capable of identifying useful ENSO indexes when adopted on different hydrological modelling contexts. A comparative analysis is conducted with respect to two traditionally adopted correlation analysis and statistical test methods.

2 IVS FOR ENSO DETECTION

IVS is an important step in environmental and water resources systems modelling. The problem of IVS arises every time one wants to model the relationship between a variable of interest and a subset of potential explanatory input variables, but there is uncertainty about which subset to use among a large number of candidate sets available [Galelli and Castelletti, 2013b]. Especially when dealing with data-driven models of hydro-climate processes, we are often faced with the dual challenge of preprocessing large sets of candidate inputs and characterizing their highly non-linear relationship to the output of interest. In this context, IVS allows the identification of the subset of input variables that, collectively, possess the largest amount of information about the system being modelled. Hence, it reduces model complexity and enhances predictive accuracy by avoiding the interference of not relevant or redundant information.

As far as the authors know, IVS algorithms have never been adopted as an ENSO detection method, but they are believed to be a potentially valuable tool with some interesting additional merits wrt the commonly adopted methods. Specifically, IVS i) provides useful quantitative information about the forecast potential of both meteorological variables and ENSO indexes; ii) as a by-product, IVS produces a data-driven model that can be used for operational purposes.

Among the different IVS algorithms available for hydrological modelling problems (see, for example, May et al. [2008]), the one adopted in this study is the Iterative Input variable Selection (IIS) algorithm, recently introduced by Galelli and Castelletti [2013b]. IIS is a hybrid model-based/model-free algorithm, which scales well to large datasets and accounts for non-linear dependencies and redundancy between the input variables. The algorithm adopts a forward selection strategy, i.e. one variable is added at each iteration. The selection process is terminated when the selection of a further input does not improve the accuracy, measured in terms of coefficient of determination R^2, of an underlying data-driven model. In this study, the IIS algorithm is combined with Extremely Randomized Trees

(Extra-Trees), a non-parametric tree-based ensemble regression method proposed by Geurts et al. [2006], which was empirically demonstrated to outperform other regression methods in terms of modelling flexibility, computational efficiency, and scalability with respect to the input dimensionality. An application of IIS to hydrological problems is presented in Galelli and Castelletti [2013a].

3 CASE STUDIES

The IVS-based procedure is evaluated on two case study, namely the Columbia River, located in the western US, and the Williams River, in eastern Australia. As the Pacific Ocean is the area where ENSO phases originate, the west coast of America and the east coast of Australia are known to be sensitive to ENSO fluctuations. Besides, the influence of ENSO on the Columbia and Williams Rivers is well-documented in the literature, and large datasets are available for both basins.

3.1 Columbia River basin

The Columbia River is the largest river in the Pacific North West (PNW) region of North America. It drains an area of approximately 669,000 km^2 at the border between the State of Washington (US) and British Columbia (Canada). The average flow measured at The Dalles (US) is approximately 5,400 m^3/s. It is primarily a snowmelt-driven system: most of the runoff volume occurs during May through July. ENSO influences streamflow in the PNW mostly during the winter period: warm-phase ENSO (El Niño) is associated with above average temperature and below average precipitation in winter months, with an increased likelihood of below-average streamflow in spring and summer; Cold-phase ENSO (La Niña) is associated with below average temperatures and above average precipitation in winter that lead to above average streamflow in spring and summer [Hamlet and Lettenmaier, 2000]. The time period considered for this case study spans over the period 1950-2000. Meteorological monthly data are obtained from the NOAA National Climatic Data Center, while naturalised streamflow data at The Dalles are obtained from the Bonneville Power Administration.

3.2 Williams River basin

The Williams River basin is located in the Hunter Region of New South Wales, Australia. It is approximately 1,300 km^2 in area (about 500 times smaller than the Columbia River basin) and the average discharge at Glen Martin is about 11 m^3/s. The system is characterised by total absence of snowfall and by extreme inter-annual variability in rainfall. As a consequence, streamflows are also highly variable from year to year with the annual flow in wet years approximately twice as much as the flow during dry years. Severe droughts can last for many months but also exceptional streamflow peaks are not uncommon. ENSO impacts the area with particularly dry conditions (lower than average rainfall and streamflow) during El Niño and very wet conditions during La Niña [Kiem and Franks, 2001]. The time period considered for this case study is 1950-2003: meteorological monthly data are obtained from the Australian Government Bureau of Meteorology, while streamflow data at Glen Martin are obtained from the NSW Government WaterInfo website.

4 RESULTS

For each case study we first adopted standard statistical analysis techniques for ENSO detection (i.e. correlation analysis and statistical tests), and then confronted the results against the IIS algorithm outcomes. In both the case studies, the output variable of interest is the streamflow. This because ENSO effects are often stronger on streamflows rather than on precipitation, since the streamflow is an integrator of the processes occurring in a river basin [Wooldridge et al., 2001].

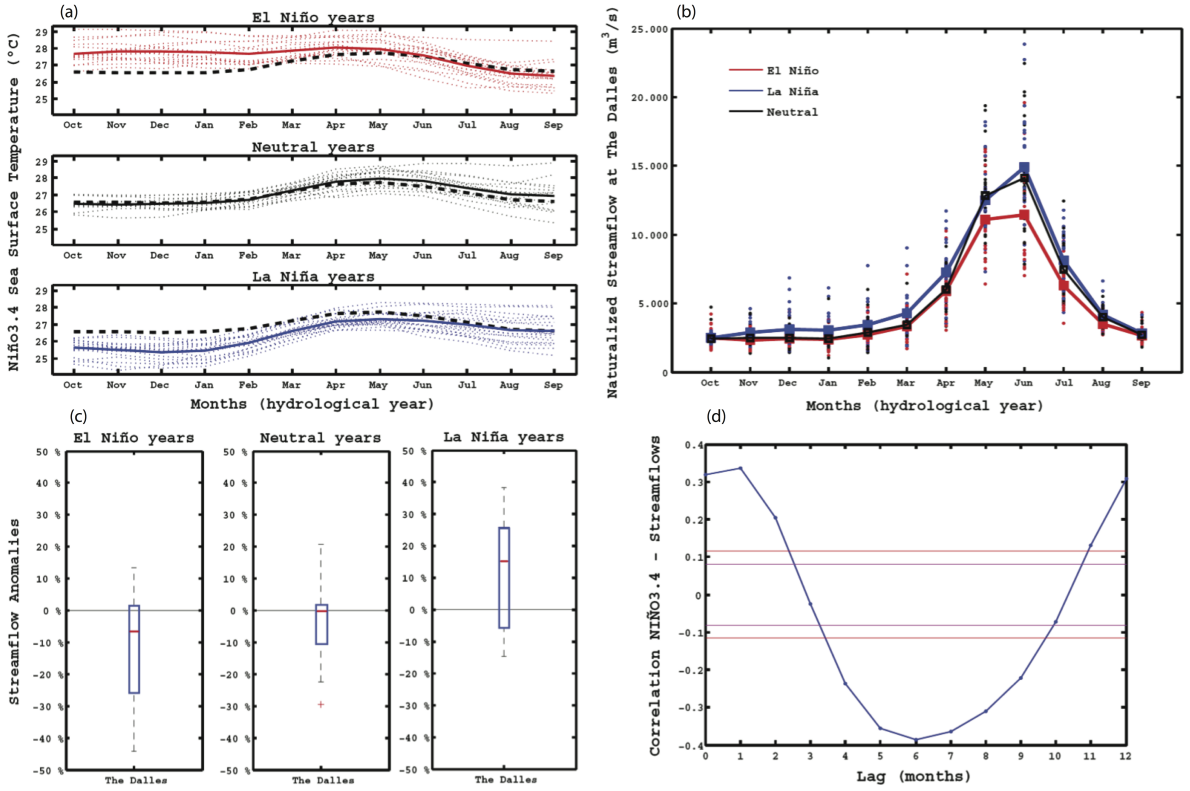

Figure 1. Graphical/statistical results for the Columbia River case study. a) ENSO signal detection; b) Hydrological response to ENSO signal; c) Box plots; d) Lag detection.

4.1 Columbia River basin

Similarly to Hamlet and Lettenmaier [2000], the ENSO index used in this study is the Niño 3.4 SST. The values of this index for each ENSO category are shown in Figure 1a: dotted lines represent single years, solid lines are average values, and the dashed black line is the long-term average over the entire period. It can be noticed that during neutral years the Niño 3.4 values spread both above and below the long-term average, whereas during El Niño years a clear trend above average exists (and conversely for La Niña years). The ENSO-hydrology relationship is confirmed by the statistical analysis developed in terms of box plots (Figure 1c): the difference within discharge anomalies during El Niño and La Niña is evident considering not only the median values, but also the variances of the distributions, suggesting that ENSO can significantly influence the overall variability of annual volumes. Finally, Figure 1d shows that the highest correlation between the Niño 3.4 index and the streamflow occurs at lag 6, confirming that the correlation takes place via snowpack, with the streamflow peak in early June being influenced by the Niño 3.4 measured in December-February. These different trends are found to reflect onto streamflows (Figure 1b): streamflows during El Niño are, in fact, visibly lower with respect to La Niña, especially during the summer period (May-July).

The IIS algorithm is employed to identify the most significant input variables for predicting the monthly streamflow at time t. The set of candidate input variables for the IVS experiment consists of both hydro-meteorological variables and ENSO indexes, for a total of 35 candidates: past rainfall $\mathbf{p}^i_{\mathbf{t-1}}$ and $\mathbf{p}^i_{\mathbf{t-2}}$ in the *i*=1,...,10 stations available, a snowmelt proxy index ($\mathbf{snowmelt}^j$) in *j*=1,...,9 stations, and the Niño 3.4 SST considered as average value over a 3- and 5-month running mean window ($\mathbf{SST3.4^{3mrm}}$ and $\mathbf{SST3.4^{5mrm}}$). We considered t-6, t-8, t-10 and t-12 time-lag for $\mathbf{SST3.4^{3mrm}}$, and t-7 and t-11 for $\mathbf{SST3.4^{5mrm}}$ (in order to avoid any overlap).

Table 1 reports the results of ten IVS experiments in terms of frequency with which a variable has

IIS			
FEATURE	FREQ	AVG RANK	% R^2
$\mathbf{snowmelt}^{\text{GOLDEN}}$	10	1.0	70.32
$\mathbf{snowmelt}^{\text{POCATELLO}}$	10	3.2	13.58
$\mathbf{SST3.4^{3mrm}_{t-6}}$	10	3.5	2.43
$\mathbf{p}^{\text{MISSOULA}}_{\mathbf{t-1}}$	10	5.3	0.51
$\mathbf{snowmelt}^{\text{BOISE}}$	9	3.8	9.55

Table 1. Results obtained with the IIS algorithm for the Columbia case study

been selected, average position during the stepwise selection process and relative contribution (in terms of R^2) to the underlying data-driven model. From Table 1 it can be seen that a high percentage of streamflow variability can be described by the snowmelt process. In particular, the snowmelt measured at Golden station (Canada) results the main driver of streamflow variability, followed by snowmelt at Pocatello (Idaho). Interestingly, the Niño 3.4 SST also results to have an impact on streamflows. Although this contribution is weaker than the snowmelt one, the variable $\mathbf{SST3.4^{3mrm}_{t-6}}$ is always selected by the IIS algorithm and it contributes to the predictive model by about 2.5 %. This result is consistent with the outcomes of the statistical analyses described above, as well as with the findings of Hamlet and Lettenmaier [2000]. In conclusion, the IVS experiments provide an insight of the main hydro-meteorological processes that take place in the basin (melting process of the snow cumulating in the mountainous regions) and selects the most significant predictors of streamflow variability among an initial set of 35 candidates, confirming the importance of ENSO at the basin scale.

4.2 Williams River basin

Figure 2 shows the results of the graphical/statistical analyses developed for the Williams River basin. In this case, the ENSO index we use is the Multivariate ENSO Index (MEI), based on Kiem and Franks [2001]. Observing the historical observations of this index (Figure 2a), it can be seen that during El Niño a clear positive trend exists (conversely for La Niña), while during neutral years the MEI values spread both above and below zero. This results in different behaviours of streamflow under the two ENSO phases (Figure 2b): streamflow during El Niño are very low, especially during summer-autumn months, whereas streamflow during La Niña are higher in almost every month of the year. Box plots representing annual streamflow anomalies are presented in Figure 2c: on one side, anomalies result strongly negative during El Niño, with high frequency of severe droughts; on the other side, during La Niña, the overall distribution visibly shifts into the positive range, and the variance is enhanced to exceptional levels, with anomalies going from -100% to almost +250%. Finally, Figure 2d presents the result of the correlation analysis between streamflow and MEI: the best correlation (about -0.15) is found to occur at a lag of 2-3 months, which is consistent with Chiew et al. [1998].

The set of candidate variables for the IVS experiments consists of both meteorological and ENSO-related variables, for a total of 10 candidates that are used to predict the streamflow at time t: past rainfall $\mathbf{p}^{i}_{\mathbf{t-1}}$ in *i*=1,...,4 stations, maximum and minimum temperature values ($\mathbf{T}^{\text{MAX}}_{\mathbf{t-1}}$ and $\mathbf{T}^{\text{MIN}}_{\mathbf{t-1}}$) measured at the station of Williamtown, and two ENSO indexes, both calculated as 3-month running mean window. These latter are $\mathbf{MEI^{3mrm}}$ [Kiem and Franks, 2001] and $\mathbf{SOI^{3mrm}}$ [Chiew et al., 1998]. With respect to these two ENSO indexes we explore lags up to 6 months, avoiding overlapping variables as much as possible, which leads to considering lag 2 and lag 5, as we are testing them in a 3-month running mean window.

Table 2 reports the results of IVS experiments. It can be seen that the IIS algorithm selects the $\mathbf{MEI^{3mrm}}$ centred to lag 2 ten times out of ten, mostly in second position. This means that there is an ENSO signal on the area, and that ENSO effectively impacts the streamflow in the Williams River basin. Moreover, MEI outperforms SOI. These results are consistent with the findings by Kiem and Franks [2001], who show that MEI outperforms other ENSO indexes in discriminating runoff variability

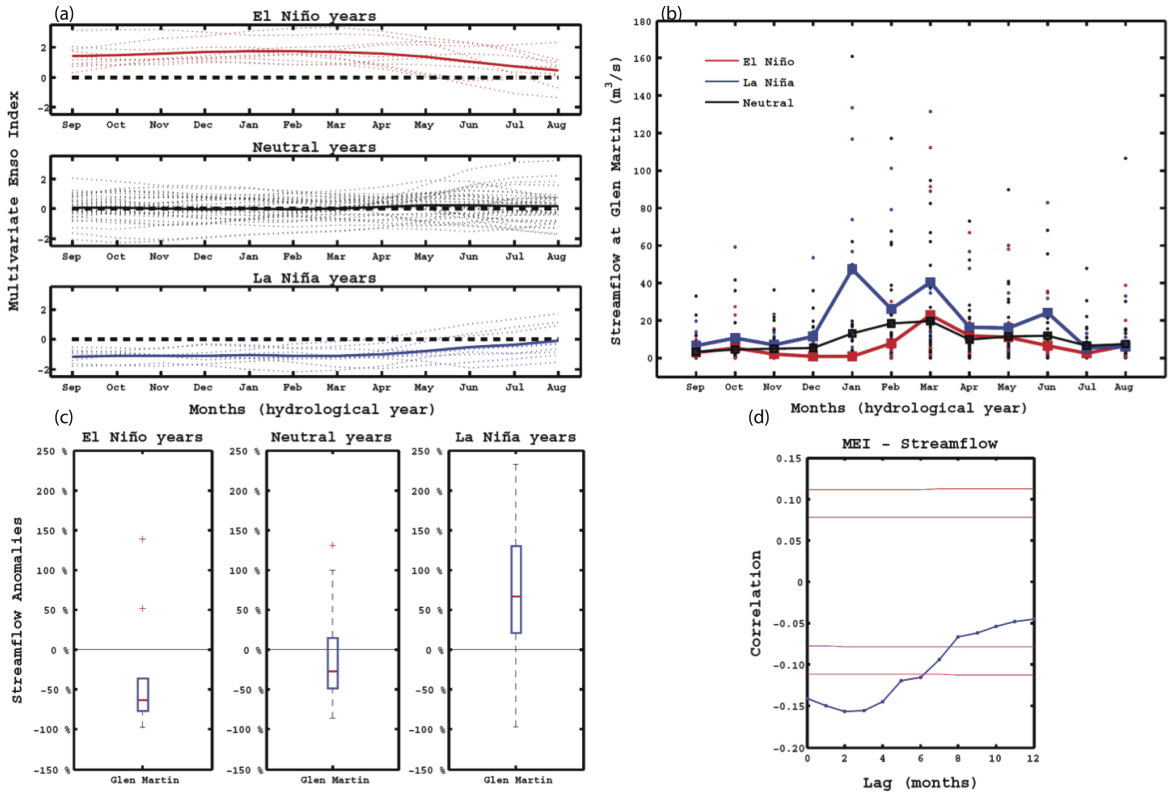

Figure 2. Graphical/statistical results for the Williams River case study. a) ENSO signal detection; b) Hydrological response to ENSO signal; c) Box plots; d) Lag detection.

IIS			
FEATURE	FREQ	AVG RANK	% R^2
$\mathbf{T}_{\mathbf{t-1}}^{\mathrm{MIN}}$	10	1.3	42.06
$\mathbf{MEI}_{\mathbf{t-2}}^{\mathbf{3mrm}}$	10	2.2	30.71
$\mathbf{p}_{\mathbf{t-1}}^{\mathrm{CHICHESTER}}$	10	2.5	15.07

Table 2. Results obtained with the IIS algorithm for the Williams case study.

in the same watershed, and by Chiew et al. [1998], who find a forecast potential for eastern Australia in ENSO-related information averaged over 2-3 months and with lag up to 3 months. Furthermore, they are in agreement with the outcomes of our previous graphical and statistical analyses.

5 Conclusions

This paper proposes a novel procedure based on IVS for quantifying ENSO impacts at the basin scale and assessing the potential for ENSO to inform medium-to-long range streamflow prediction models. The IVS-based procedure is compared to some of the most commonly employed ENSO detection methods, and evaluated on two real-world monthly streamflow prediction problems in the PNW region of the US and in South-East Australia. The results show that IVS outcomes are consistent with those from traditional ENSO detection methods and in agreement with what is known from previous studies conducted in the same regions. This is true for both case studies, meaning that IVS is capable of assessing ENSO-related effects on different hydrological modelling contexts. Further research will focus on a comparison between the IIS and other IVS algorithms.

ACKNOWLEDGMENTS

This study work was completed while Ludovica Beltrame and Daniele Carbonin were on leave at the Singapore University of Technology and Design with the financial support of Politecnico di Milano. The third author is supported by the SRG ESD 2013 061 Start-up Research project.

REFERENCES

Chiew, F. H. S. and McMahon, T. A. (2002). Global enso-streamflow teleconnection, streamflow forecasting and interannual variability. *Hydrological Sciences Journal*, 47(3):505–522.

Chiew, F. H. S., Piechota, T. C., Dracup, J. A., and McMahon, T. A. (1998). El niño/southern oscillation and australian rainfall, streamflow and drought: Links and potential for forecasting. *Journal of Hydrology*, 204(1-4):138–149.

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

Galelli, S. and Castelletti, A. (2013b). Tree-based iterative input variable selection for hydrological modeling. *Water Resources Research*, 49(7):4295–4310.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extreme randomized trees. *Machine learning*, 63(1):3–42.

Gutierrez, F. and Dracup, J. A. (2001). An analysis of the feasibility of long-range streamflow forecasting for colombia using el niño southern oscillation indicators. *Journal of Hydrology*, 246(1-4):181–196.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hamlet, A. F. and Lettenmaier, D. P. (2000). Long-range climate forecasting and its use for water management in the pacific northwest region of north america. *Journal of Hydroinformatics*, 2(3):163–182.

Hidalgo, H. G. and Dracup, J. A. (2003). Enso and pdo effects on hydroclimatic variations of the upper colorado river basin. *Journal of Hydrometeorology*, 4(1):5–23.

Kahya, E. and Dracup, J. A. (1993). U.s. streamflow patterns in relation to the el niño/southern oscillation. *Water Resources Research*, 29(8):2491–2503.

Kiem, A. S. and Franks, S. W. (2001). On the identification of enso-induced rainfall and runoff variability: a comparison of methods and indices. *Hydrological Sciences Journal*, 46(5):715–727.

May, R. J., Maier, H. R., Dandy, G. C., and Fernando, T. M. K. (2008). Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling & Software*, 23(10-11):1312–1326.

Opitz-Stapleton, S., Gangopadhyay, S., and Rajagopalan, B. (2007). Generating streamflow forecasts for the yakima river basin using large-scale climate predictors. *Journal of Hydrology*, 341(3-4):131–143.

Ropelewski, C. F. and Halpert, M. S. (1986). North american precipitation and temperature patterns associated with the el niño/southern oscillation (enso). *Monthly Weather Review*, 114(12):2352–2362.

Sharma, A. (2000). Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - a strategy for system predictor identification. *Journal of Hydrology*, 239(1-4):232–239.

Simpson, H. J., Cane, M. A., Herczeg, A. L., and Zebiak, S. E. (1993). Annual river discharge in southeastern australia related to el niño-southern oscillation forecasts of sea surface temperatures. *Water Resources Research*, 29(11):3671–3680.

Wooldridge, S. A., Franks, S. W., and Kalma, J. D. (2001). Hydrological implications of the southern oscillation: variability of the rainfall-runoff relationship. *Hydrological Sciences Journal*, 46(1):73–88.