Journal of Environmental Science International
22(8); 953~963; August 2013

pISSN: 1225-4517 eISSN: 2287-3503
http://dx.doi.org/10.5322/JESI.2013.22.8.953

ORIGINAL ARTICLE

다중선형회귀분석에 의한 계절별 저수지 유입량 예측

강 재 원*
충남대학교 국제수자원연구소

Forecasting of Seasonal Inflow to Reservoir Using Multiple Linear Regression

Jaewon Kang*
International Water Resources Research Institute, Chungnam National University, Daejeon 305-764, Korea

Abstract

Reliable long-term streamflow forecasting is invaluable for water resource planning and management which allocates water supply according to the demand of water users. Forecasting of seasonal inflow to Andong dam is performed and assessed using statistical methods based on hydrometeorological data. Predictors which is used to forecast seasonal inflow to Andong dam are selected from southern oscillation index, sea surface temperature, and 500 hPa geopotential height data in northern hemisphere. Predictors are selected by the following procedure. Primary predictors sets are obtained, and then final predictors are determined from the sets. The primary predictor sets for each season are identified using cross correlation and mutual information. The final predictors are identified using partial cross correlation and partial mutual information. In each season, there are three selected predictors. The values are determined using bootstrapping technique considering a specific significance level for predictor selection. Seasonal inflow forecasting is performed by multiple linear regression analysis using the selected predictors for each season, and the results of forecast using cross validation are assessed. Multiple linear regression analysis is performed using SAS. The results of multiple linear regression analysis are assessed by mean squared error and mean absolute error. And contingency table is established and assessed by Heidke skill score. The assessment reveals that the forecasts by multiple linear regression analysis are better than the reference forecasts.

Key words : Forecasting seasonal inflow, Sea surface temperature, Geopotential height, Partial mutual information, Heidke skill score

1. 서 론

신뢰성 있는 장기간 하천유출량 예측은 물 사용자들의 수요에 반응하여 물공급을 할당해야 하는 수자원 계획과 관리 측면에서 매우 중요하다. 우리나라의 경우 계절별 하천유출량이 매우 변화가 커서 농업계획과 다른 물관리 목적들에 상당한 어려움을 부과하고 있다. 유출량의 자연적 변동성으로 인해 이용 가능한 수량이 바라는 것보다 훨씬 적거나 많은 경우가 빈번히 발생하고 있다. 가뭄은 물공급과 식량의 수급 등에 매우 심각한 영향을 미치게 된다. 그리고 가뭄은 소비적인 사용을 위한 이용 가능량의 감소이외에도 가

Received 28 December, 2012; **Revised** 20 March, 2013;
Accepted 17 April, 2013
***Corresponding author** : Jaewon Kang, International Water Resources Research Institute, Chungnam National University, Daejeon 305-764, Korea
Phone: +82-42821-8958
E-mail: zipthin@hanmail.net

묶기간 동안의 저수량(low flow) 감소는 수질의 악화, 전력생산량 감소, 어류와 야생동물에 대한 영향 등을 야기할 수 있다.

장기간 하천유출량을 예측하기 위해서 통계적인 모형을 이용하는 경우 먼저 적절한 예측인자의 선정이 이루어져야 한다. 유출량의 유용한 예측인자를 식별한 것은 장기예측모형의 중요한 단계이다. 예측인자의 선정에 가장 널리 이용되는 방법은 상관계수를 이용하는 것이다. 단변량자료에 대해서 시계열분석을 수행하는 경우에는 자기 상관계수와 부분 자기상관계수를 이용하여 모형을 식별하고 이를 통해 예측을 수행할 수 있다(Hipel과 McLeod, 1994; Wei, 1990). 다중선형회귀분석의 예측인자를 선정하는 경우에는 교차상관계수가 이용되며, 여러 개의 잠정적인 예측인자들 중에서 선택해야 하는 경우에는 부분 교차상관계수가 이용된다(Hipel과 McLeod, 1994). Sharma(2000)는 상관계수가 선형적인 관계만을 나타내기 때문에 비선형을 나타내는 복잡한 시스템에서는 잘못된 결과를 줄 수 있으므로 정보이론의 상호정보량(mutual information)을 이용하여 예측인자를 식별할 것을 제안하였다. 그리고 상호정보량만으로는 여러 개의 예측인자들이 존재할 경우에 이미 존재하는 예측인자를 조건으로 하여 다음 예측인자를 식별할 수 없음을 지적하고 부분 상호정보량(partial mutual information)을 제안하였다. 부분 상호정보량은 원래의 자료에서 이미 식별된 예측인자의 조건부평균을 뺀 잔차들 간의 상호정보량을 계산하는 것으로서 부분 교차상관계수를 구하는 과정과 유사하다. Sharma 등(2000)은 이러한 방법을 이용하여 호주의 한 유역에 대한 강수량의 예측인자를 식별하였다.

Chiew 등(1998)은 El Niño/Southern Oscillation (ENSO)과 Australia의 강우, 유출량, 가뭄 등과의 관계를 검토한 결과 대부분의 지역에서 통계적으로 유의함을 나타냈지만 정확한 예측을 수행할 수 있을 정도로 강한 일관성을 보이지 않음을 밝혔다. 그러나 원격상관(teleconnection)이 하반기에 더 강한 점을 이용하여 일부 지역의 봄과 여름의 강우량을 예측하는 데 ENSO를 이용할 수 있음을 제안하였다. 그리고 유출량의 경우에는 ENSO와의 지체 상관보다는 계열상관이 더 크게 나타남으로 유출량을 예측할 때에 두 가지를 같이 고려해야 함을 제안하였다. 그리고 Simpson과 Colodner(1999)는 엘리뇨 정보가 수자원관리 기술을 향상시킬 수 있는 잠재성에 긍정적인 평가를 했으며, Liu 등(1998)은 ENSO 예측의 포함여부에 따른 수문학적 모형들의 예측능력을 비교하고 ENSO 관측과 예측을 통합함에 의해서 모형의 예측능력이 상당히 개선됨을 발견했다. Berri와 Flamenco(1999)는 엘리뇨 관측값과 예측값을 사용해서 단계적 다중선형회귀(stepwise multiple linear regression) 방법으로 모형을 구성하고 계절별 유출량 예측을 실시하여 엘리뇨 예측값의 사용가능성을 보였다. 10월~3월 유출용적(연유출의 70% 차지)과 NINO3($5^\circ \times 5^\circ$, $150^\circ W \times 90^\circ W$ 격자의 산술평균) 지역의 해수면온도(Sea Surface Temperature, SST)의 상관관계를 구한 결과 이전기간 3월~4월, 동일기간 11월~12월과 양의 상관관계가 있는 것으로 판명됨에 따라 11월~12월의 해수면온도는 6개월 전(5월)에 예측된 값을 이용하였다.

해수면온도와 같은 전 지구적인 기상자료를 이용하여 수문자료 예측을 시도한 연구자들의 적용지역을 살펴보면 다음과 같다. Awadallah와 Rousselle(2000)은 High Aswan Dam의 상류 지점의 여름철 유출량을 예측하는 데 이용하였으며, Uvo와 Graham(1998)은 northern South America에서 13지역, 아마존 유역의 6개 지역을 대상으로 하였다. Berri와 Flamenco(1999)는 Argentina의 Andes 산맥 중앙부에 있는 Diamante 강의 La Jaula에서 측정된 계절별 유출량(유역면적 2,750 km^2)을 이용하였으며, Simpson 등(1993)은 호주 남동부의 Murray와 Darling 강을 대상으로 하였다. Piechota 등(1998)은 호주 동부 10개 지점 자료를 이용하였으며, Piechota 등(2001)은 호주 5개 지점 자료를 대상으로 분석하였다. Sharma 등(2000)은 호주 Sydney 근방의 Warragamba 댐의 계절별 강우량을 이용하였으며, Ahn과 Park(2000)은 서울, 부산, 광주, 대구 지방의 월별 기온과 강수량을 대상으로 예측을 실시하였다.

기존의 연구에서 식별된 예측인자의 지체에 대해서 살펴보면 다음과 같다. Awadallah와 Rousselle (2000)은 0~24개월, Sharma 등(2000)은 2년 정도의 지체를 이용하였다. 그리고 Ahn과 Park(2000)은 0~12개월 지연시켜 상관관계를 분석(월자료 대상)하였

으며, Piechota 등(1998)은 직전 계절의 자료를 이용하였다. Simpson 등(1993)은 지체 1년 정도, Berri와 Flamenco(1999)는 1~2년 전의 유출량을 예측인자로 사용(상관관계로부터 도출)하였으며, Uvo와 Graham (1998)은 한 계절 전의 해수면온도를 이용하였다.

본 연구에서는 안동댐 계절별 유입량을 예측하기 위해서 남방진동지수(Southern Oscillation Index, SOI), 해수면온도(Sea Surface Temperature, SST), 500 hPa 지위고도(Geopotential Height, GPH) 자료로부터 예측인자를 선정하였다. 선정된 인자를 사용하여 통계적 방법인 다중선형회귀분석으로 안동댐 계절별 유입량 예측을 수행하고, 회귀모형의 예측능력을 교차확인(cross validation)을 통해서 평가하였다.

2. 자료 및 방법

2.1. 예측인자 선정방법

2.1.1. 상관계수에 의한 예측인자 선정 방법

두 변수간의 선형 관계를 나타내는 공분산은 그 크기가 변수의 단위나 범위에 영향을 받기 때문에 공분산과 분산의 추정값을 대입해서 얻은 식 (1)의 교차상관계수를 예측인자 선정에 이용하였다.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

여기서 $\bar{x}$, $\bar{y}$는 표본평균을 의미한다.

여러 개의 예측인자가 식별된 경우에 예측변수에 영향을 미치는 인자를 선별하는 경우에 유의하다고 판단된 예측인자를 조건으로 하여 이와 유사한 영향을 나타내는 예측인자를 제거하기 위한 방법으로 식 (2)의 부분 교차상관계수를 이용하였다. 예를 들어 예측변수 Y에 대해서 잠정적인 예측인자 X_1, X_2가 있다고 하면 다음 식 (2)와 같이 예측인자 X_1을 조건으로 하는 부분 교차상관계수를 계산하여 X_2를 예측인자로 사용할 것인지를 결정하였다.

$$r_{YX_2,X_1} = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{\sqrt{(1-r_{YX_1}^2)(1-r_{X_1X_2}^2)}} \quad (2)$$

2.1.2. 상호정보량을 이용한 예측인자 선정 방법

엔트로피(entropy)는 확률변수(random variable)의 불확실성(uncertainty)의 척도이다. 즉 확률변수를 기술하기 위해서 평균적으로 필요한 정보량의 척도로서 연속 확률변수의 엔트로피 $h(X)$는 다음 식 (3)과 같이 정의되며, $f(x)$는 확률밀도함수를 나타낸다(Cover와 Thomas, 1991).

$$h(X) = -\int_s f(x)\log f(x)dx \quad (3)$$

연속적인 확률변수의 상호정보량(mutual information; MI)은 다음 식 (4)와 같이 정의되며, $f(x,y)$는 결합확률밀도함수이고 $f(x)$, $f(y)$는 X, Y의 확률밀도함수이다(Cover와 Thomas, 1991). 상호정보량은 두 변수의 연관성을 나타낸다.

$$MI(X;Y) = \iint f(x,y)\ln\left[\frac{f(x,y)}{f(x)f(y)}\right]dxdy \quad (4)$$

다중 예측인자의 식별은 예측변수와 예측인자 집합사이의 부분 종속성(partial dependence)에 대한 척도(measure)가 필요하다. Sharma(2000a)는 MI 판단기준이 자료사이의 부분 종속성의 척도로서 직접 사용될 수 없음을 지적하고 식 (5)와 같은 부분 상호정보량(partial mutual information; PMI)의 사용을 제안했다. 이미 존재하는 예측인자들의 집합 $\mathbf{z}$에 대해서 y와 X사이의 PMI는 다음 식 (5)와 같이 표현된다.

$$PMI = \iint f(\tilde{x},\tilde{y})\ln\left[\frac{f(\tilde{x},\tilde{y})}{f(\tilde{x})f(\tilde{y})}\right]d\tilde{x}d\tilde{y} \quad (5)$$

여기서 $\tilde{x} = x - E[X \mid \mathbf{z}]$, $\tilde{y} = y - E[Y \mid \mathbf{z}]$이고, 원 자료에서 식 (6)으로 표현되는 조건부 기대값을 뺀 $\tilde{x}$, $\tilde{y}$가 기존의 예측인자 $\mathbf{z}$의 영향을 고려한 x와 y의 잔차정보(residual information)를 나타낸다.

$$E[X \mid \mathbf{z}] = \sum_{i=1}^{n} w_i\left\{x_i + (\mathbf{z}-\mathbf{z}_i)'\mathbf{S}_{zz}^{-1}\mathbf{S}_{xz}\right\} \quad (6)$$

여기서 $\mathbf{S}_{xz}$는 표본 교차공분산이고 $\mathbf{S}_{zz}$는 표본 공분산이다. 그리고 가중값 w_i는 다음 식 (7)과 같고 h는 평활모수이다.

$$w_i = \frac{\exp\left(-\frac{(\mathbf{z}-\mathbf{z}_i)'\mathbf{S}_{zz}^{-1}(\mathbf{z}-\mathbf{z}_i)}{2h^2}\right)}{\sum_{j=1}^{n}\exp\left(-\frac{(\mathbf{z}-\mathbf{z}_j)'\mathbf{S}^{-1}{}_{zz}(\mathbf{z}-\mathbf{z}_j)}{2h^2}\right)} \quad (7)$$

(부분) 교차상관계수의 경우에는 유의수준에 따른 임계값을 설정할 수 있으나 (부분) 상호정보량의 경우에는 수식으로 표현된 유의수준에 따른 임계값이 존재하지 않는다. 따라서 (부분) 상호정보량과 (부분) 교차상관계수의 상호보완적 사용을 원활하게 하기 위해서 유의수준을 설정하는 데 붓스트랩 방법을 사용하였다. 본 연구에서 잠재적인 예측인자를 선정하는 경우에 교차상관계수와 상호정보량이 모두 유의수준 5%에서 유의한 경우에 한하였다. 상호정보량의 경우 선형적인 관계가 없이 자료가 분포된 경우에도 큰 값을 나타낼 수 있으며, 교차상관계수의 경우에는 대부분의 자료와 상당히 떨어진 자료 값이 존재하는 경우 이에 대한 영향을 심각하게 받는 경우가 있기 때문이다. 그리고 상호정보량의 경우에는 양의 상관관계나 음의 상관관계에 대한 정보를 전혀 제공하지 않는다.

2.1.3. 붓스트랩(Bootstrapping)

붓스트랩 방법은 자료 집합 $\mathbf{x}$로부터 붓스트랩 표본 $\mathbf{x}^*$를 균등난수를 이용하여 생성하고 이로부터 대상으로 하는 모수에 대한 붓스트랩 재추정값들(bootstrap replications)을 구하여 원자료로부터 구한 모수의 추정값에 대한 신뢰구간 등의 통계적 추론을 위하여 개발된 것이다(Efron과 Tibshirani, 1993). 붓스트랩 표본 $\mathbf{x}^* = (x_1^*,\ x_2^*,\ \cdots, x_n^*)$는 원래의 자료 $x_1, x_2, \cdots, x_n$으로부터 임의로 n번 복원추출해서 구성되고, 이 과정을 B번 시행하면 각각의 크기가 n인 $\mathbf{x}^{*1}$, $\mathbf{x}^{*2}$, $\cdots$, $\mathbf{x}^{*\mathrm{B}}$인 표본을 얻을 수 있으며, 이로부터 붓스트랩 재추정값 $s(\mathbf{x}^{*\mathrm{b}})$를 구하는 과정을 Fig. 1에 나타내었다. 본 연구에서 붓스트랩 재추정값은 (부분) 교차상관계수와 (부분) 상호정보량을 붓스트랩 표본으로부터 계산한 것을 의미한다.

붓스트랩 백분위수에 기초한 신뢰구간은 Fig. 1과 같이 붓스트랩 자료 집합 $\mathbf{x}^*$를 생성하고 이로부터 붓스트랩 재추정값 $\hat{\theta}^* = s(\mathbf{x}^*)$를 계산하여 구하였다. $\hat{G}$을 $\hat{\theta}^*$의 누가분포함수라 하면 유의수준 α에 대한 $1-2\alpha$ 백분위수 구간은 $\hat{G}$의 α와 $1-\alpha$ 백분위수들에 의해 다음 식 (8)과 같이 표현할 수 있다(Efron과 Tibshirani, 1993).

$$\left[\widehat{\theta_{\%,lo}},\ \widehat{\theta_{\%,up}}\right] = \left[\widehat{G^{-1}}(\alpha),\ \widehat{G^{-1}}(1-\alpha)\right] \quad (8)$$

식 (8)에 의해서 붓스트랩 분포의 $100\cdot\alpha^{th}$ 백분위수는 $\widehat{G^{-1}}(\alpha) = \widehat{\theta^{*(\alpha)}}$이므로 백분위수 구간을 다음 식 (9)와 같이 쓸 수 있다.

$$\left[\widehat{\theta_{lo}},\ \widehat{\theta_{up}}\right] = \left[\widehat{\theta^{*(\alpha)}},\ \widehat{\theta^{*(1-\alpha)}}\right] \quad (9)$$

식 (8)과 (9)는 붓스트랩 재추정값들이 무한대인 이상적인 붓스트랩 상황에 관한 것이지만, 본 연구에서는 유한한 B개의 추정값들을 이용하였다. B개의 독립적인 붓스트랩 자료 집합 $(\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*\mathrm{B}})$를 생성하고 붓스트랩 재추정값 $\hat{\theta}^*(b) = s(\mathbf{x}^{*\mathrm{b}})$, $\mathrm{b} = 1,2,\cdots,\mathrm{B}$를 계산하였다. $\theta_B^{*(\alpha)}$를 $\hat{\theta}^*(b)$의 $100\cdot\alpha^{th}$ 경험적 백분위수 즉 $\hat{\theta}$에 대한 B개의 추정값들의 순서화된 자료에서 $B\cdot\alpha^{th}$ 값이라 하고, 유사하게 $\theta_B^{*(1-\alpha)}$를 $\hat{\theta}^*(b)$의 $100\cdot(1-\alpha)^{th}$ 경험적 백분위수라 하면, 근사적인 $1-2\alpha$ 백분위수 구간은 다음 식 (10)과 같이 표현할 수 있다.

$$\left[\widehat{\theta_{lo}},\ \widehat{\theta_{up}}\right] \approx \left[\widehat{\theta_B^{*(\alpha)}},\ \widehat{\theta_B^{*(1-\alpha)}}\right] \quad (10)$$

예를 들어 $B = 2{,}000$, $\alpha = 0.05$라면 백분위수 구간 $\left(\theta^{*(0.05)},\ \theta^{*(0.95)}\right)$는 2,000개의 $\hat{\theta}^*(b)$의 순서화된 값들 중에서 $(100^{th},\ 1{,}900^{th})$가 된다.

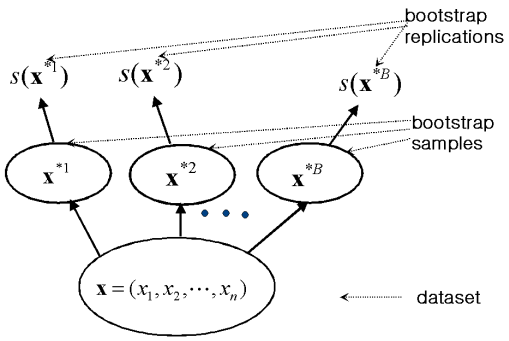

Fig. 1. Schematic of the bootstrap process for estimating the statistic $s(\mathbf{x})$ (Adapted from Efron and Tibshirani, 1993).

3. 결과 및 고찰

3.1. 예측인자 선정결과

3.1.1. 자료

안동댐 계절별 유입량을 예측하기 위한 예측인자의 선정에 사용한 자료는 남방진동지수(SOI), 해수면온도(SST) 그리고 북반구 500 hPa 지위고도(GPH) 자료이다. 이들 자료를 이용한 이유는 지금까지의 연구에서 이들 자료를 이용하여 장기예측을 수행한 결과 예측능력이 향상되었음을 보여주었기 때문이다(Chiew *et al*., 1998; Liu *et al*., 1998; Simpson과 Colodner, 1999). 안동댐 유입량 자료는 한국수자원공사에서 취득한 것으로 1954년 1월~1999년 12월까지의 월자료를 이용하였다. 안동댐 유역면적은 1,584 km^2이며, 1977년에 댐이 준공되었기 때문에 1977년 이전 자료는 안동댐 유역 유출량으로 표현하는 것이 타당할 수도 있으나 댐 준공이후 운영과정에서 유입량이라 명명하는 것에 따라 전체 자료를 유입량으로 표현하였다. 남방진동지수 자료는 미국 Climate Prediction Center에서 취득하였으며 1950년 1월~1999년 12월까지의 월자료를 이용하였다. 500 hPa 지위고도자료와 해수면온도자료는 서울대학교 대기과학과에서 취득한 1950년 1월~1999년 12월까지의 월자료를 사용하였다. 500 hPa 지위고도자료는 $0^\circ \sim 90^\circ N$의 북반구 자료를 이용하였으며, 위도와 경도 $2.5^\circ \times 2.5^\circ$ 격자점의 자료를 산술평균에 의한 $10^\circ \times 10^\circ$ 자료로 변환한 총 324개의 격자점을 대상으로 하여 계절별 자료에 대해서 예측인자를 식별하였다. 해수면온도자료는 $29^\circ S \sim 61^\circ N$ 사이의 자료를 이용하였으며, 위도와 경도 $2^\circ \times 2^\circ$ 격자점의 자료를 산술평균에 의해서 $10^\circ \times 10^\circ$ 자료로 변환하여 분석에 사용하였다. $2^\circ \times 2^\circ$로 관측된 자료를 $10^\circ \times 10^\circ$로 36개 지점을 평균한 총 360개의 격자점 중에서 결측이나 육지가 포함된 지역을 제외한 165개의 격자점을 대상으로 하여 계절별 자료에 대해서 예측인자를 식별하였다. 취득한 월별 자료를 이용하여 계절별 평균을 구하고 (부분) 교차상관계수와 (부분) 상호정보량을 상호보완적으로 이용하여 예측인자를 식별하였다.

3.1.2. 예측인자 선정방법의 검토

예측인자를 식별하기 위한 유의수준을 설정할 때 붓스트랩 기법을 적용하여 계산된 유의수준과 통계해석 패키지 SAS의 결과를 비교하면 Table 1과 같다. 사용한 자료는 봄철의 유입량 자료와 $(29^\circ S \sim 19^\circ S) \times (170^\circ W \sim 180^\circ W)$ 지점의 지체 3인 자료(X_1), $(29^\circ S \sim 19^\circ S) \times (90^\circ E \sim 100^\circ E)$ 지점의 지체 6인 자료(X_2), $(29^\circ S \sim 19^\circ S) \times (160^\circ W \sim 170^\circ W)$ 지점의 지체 3인 자료(X_3)와 지체 12인 자료(X_4)이다. X_1을 조건으로 하는 부분 교차상관계수와 붓스트랩의 5% 유의수준 값을 계산하고 이를 SAS 수행결과로 주어진 부분 교차상관계수와 유의수준의 확률을 이용하여 비교하였다. 작성한 프로그램의 결과와 SAS의 결과가 완벽하게 같지는 않지만 채택과 기각의 관점에서 보면 동일한 선택을 한다는 것을 볼 수 있다. 따라서 붓스트랩 기법을 이용하여 유의수준을 설정하는 방법이 SAS의 결과와 비교했을 때 타당성을 갖는다고 판단하였다.

또한 교차상관계수와 상호정보량을 이용하여 예측인자를 선정할 때 발생하는 문제점(상관계수는 크고 상호정보량은 작은 경우 또는 그 반대의 경우에 예측인자를 선정하는 문제)에 대해서 고찰하였다. Fig. 2a는 교차상관계수는 크지만 상호정보량은 작은 경우의 산포도를 나타낸 것이다(교차상관계수: 0.4022, 상호정보량: 0.1409). 겨울철 유입량 자료와 $(19^\circ S \sim 9^\circ S) \times$

Table 1. Comparison of 5% significance level using bootstrapping and SAS

Variable	Program Results		SAS Results	
	Partial Cross Correlation	The Value of 5% Significance Level	Partial Cross Correlation	The Probability of Significance Level
(X_2, Y)	0.303	0.309	0.300	0.053
(X_3, Y)	0.020	0.301	0.044	0.782
(X_4, Y)	0.263	0.293	0.280	0.072

($0^\circ \sim 10^\circ W$) 지점의 지체 7일 때 자료사이의 산포도이다. 산포도에서 보면 상당히 떨어진 몇 점에 의해서 교차상관계수가 크게 나온 것을 알 수 있다. 그리고 Fig. 2b는 교차상관계수는 작지만 상호정보량은 큰 경우를 나타낸 것이다(교차상관계수 0.0218, 상호정보량 0.2872). 여름철 유입량 자료와 ($29^\circ S \sim 19^\circ S$)×($170^\circ E \sim 180^\circ E$) 지점의 지체 1일 때 자료사이의 산포도이다. 산포도에서 보면 직선적인 관계는 거의 없으며 자료가 고르게 분포함을 볼 수 있다. 이를 근거로 예측인자를 선정할 때 교차상관계수와 상호정보량이 동시에 유의수준 5%에서 유의한 경우를 1차적으로 선정하였다. 1차적으로 선택된 자료에서 부분 교차상관계수와 부분 상호정보량을 이용하여 최종적인 예측인자를 식별하였다.

3.1.3. 예측인자의 선정결과

Table 2(a)~Table 2(d)는 최종적으로 확정된 예측인자의 속성을 나타낸 것이다. 봄철과 여름철의 Predictor 3을 제외하면 지체가 9계절 이상으로 앞에서 언급한 기존의 외국 연구결과보다 매우 긴 지체가 선정된 것을 볼 수 있다. 봄-가을, 가을-봄의 유입량 자료의 상관계수가 각각 0.35, 0.40으로 유의수준 5%에서 유의하지만, 예측인자의 선정과정에서 선택되지 않았다.

Table 2(a). Predictors for seasonal inflow forecasting: Spring

	Predictor 1	Predictor 2	Predictor 3
Classification of Predictor	GPH	GPH	GPH
Location	$60^\circ N - 70^\circ N$ $30^\circ E - 40^\circ E$	$70^\circ N - 80^\circ N$ $30^\circ W - 40^\circ W$	$60^\circ N - 70^\circ N$ $160^\circ W - 170^\circ W$
Lag (Season)	9	14	5
Cross Correlation	0.349	0.430	0.368
Mutual Information	0.203	0.165	0.152

Table 2(b). Predictors for seasonal inflow forecasting: Summer

	Predictor 1	Predictor 2	Predictor 3
Classification of Predictor	GPH	SST	SST
Location	$30^\circ N - 40^\circ N$ $60^\circ W - 70^\circ W$	$19^\circ S - 9^\circ S$ $170^\circ W - 180^\circ W$	$19^\circ S - 9^\circ S$ $160^\circ W - 170^\circ W$
Lag (Season)	9	16	1
Cross Correlation	0.425	-0.491	-0.353
Mutual Information	0.211	0.177	0.166

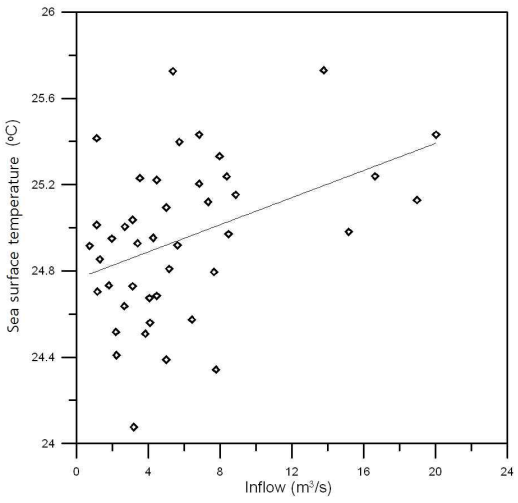

Fig. 2(a). Scatter plot of the data: representing high cross correlation and low mutual information.

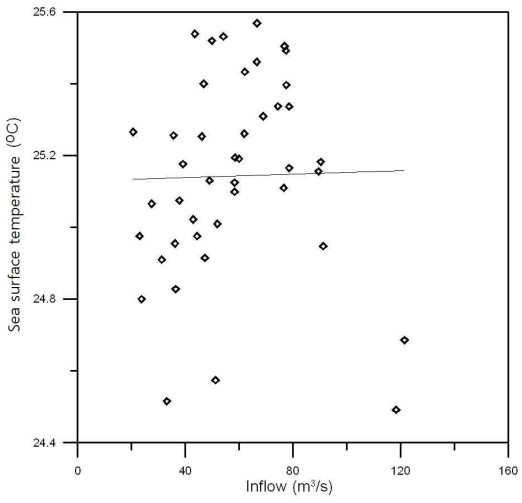

Fig. 2(b). Scatter plot of the data: representing low cross correlation and high mutual information.

Table 2(c). Predictors for seasonal inflow forecasting: Autumn

	Predictor 1	Predictor 2	Predictor 3
Classification of Predictor	SST	SST	GPH
Location	$19^\circ S$– $9^\circ S$ $50^\circ E$– $60^\circ E$	$29^\circ S$– $19^\circ S$ $120^\circ W$– $130^\circ W$	$30^\circ N$– $40^\circ N$ $110^\circ W$– $120^\circ W$
Lag (Season)	12	10	15
Cross Correlation	0.433	-0.490	0.470
Mutual Information	0.197	0.193	0.217

Table 2(d). Predictors for seasonal inflow forecasting: Winter

	Predictor 1	Predictor 2	Predictor 3
Classification of Predictor	SST	GPH	SST
Location	$19^\circ S$– $9^\circ S$ $20^\circ W$– $30^\circ W$	$40^\circ N$– $50^\circ N$ $20^\circ W$– $30^\circ W$	$9^\circ S$– $1^\circ N$ $150^\circ E$– $160^\circ E$
Lag (Season)	10	11	15
Cross Correlation	0.547	0.464	0.476
Mutual Information	0.210	0.180	0.177

3.2. 다중선형회귀분석에 의한 예측 결과 및 평가

회귀분석이란 독립변수(예측인자)를 이용하여 종속변수(반응변수)의 값을 예측하거나 독립변수가 종속변수에 미치는 영향을 측정하는 통계적 방법이다. 다중선형회귀(multiple linear regression) 모형은 다음 식 (11)과 같다(Draper와 Smith, 1998).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \,, \quad i = 1,2,\cdots,n\,(n > p) \tag{11}$$

여기서 y_i는 예측변수, β_i는 회귀계수, x_{ip}는 예측인자이다. 그리고 n은 예측인자의 관측값의 수이고 p는 예측인자의 개수이다.

분산팽창인자를 이용하여 다중공선성에 대한 검토를 수행하였으며, 그 결과를 Table 3에 나타내었다. 분산팽창인자가 10을 넘는 경우에 독립변수들 간에 다중공선성이 존재하는 것으로 판단할 수 있다(Hocking, 1996). Table 3의 결과에 의해서 분석대상으로 하는 예측인자들 사이에 다중공선성이 존재하지 않는 것으로 판단하였다. 선택된 예측인자를 이용하여 다중선형회귀모형을 적합 시킨 후 잔차가 가정을 만족하는지를 검토한 결과는 Table 4와 같다. Table 4에서 보면 잔차의 Durbin-Watson 검정으로부터 통계량이 2 근방의 값을 가지므로 독립성이 만족됨을 알 수 있으며(Table에서 괄호안의 값은 잔차의 지체 1 상관계수이다), White 검정으로부터 유의수준의 확률이 모두 0.05보다 크므로 등분산성이 만족됨을 알 수 있다(괄호안의 값은 유의수준의 확률을 나타낸다). 그러나 잔차의 정규성을 Shapiro-Wilk 검정을 통하여 검토한 결과 가을을 제외한 나머지 계절은 정규성을 만족하지 못한 것으로 나타났다(괄호안의 값은 유의수준의 확률을 나타낸다). 따라서 이들 계절의 유입량 자료에 대한 Box-Cox 변환을 실시하였으며, 변환에 사용된 계수는 0.55(봄), 0.15(여름), 0.25(겨울)이다. Box-Cox 변환한 자료를 대상으로 회귀모형을 적합 시킨 후 잔차에 대한 검정결과는 Table 5와 같다(괄호안의 값은 Table 4의 설명과 동일하다). Table 5에서 보면 변환된 자료의 잔차분석 결과 독립성, 등분산성, 정규성을 만족하는 것으로 나타났다. 적합된 다중선형회귀모형의 회귀식을 Table 6에 나타내었다. Table 6에서 가을철을 제외한 나머지 계절들은 정규성을 만족시키기 위해서 Box-Cox 변환을 수행하였으므로 역변환하여 나타낸 것이다.

Table 3. Variance inflation factor

Predictor	Spring	Summer	Autumn	Winter
x_1	1.01	1.05	1.04	1.10
x_2	1.02	1.05	1.01	1.03
x_3	1.01	1.01	1.04	1.12

Table 4. Diagnostic check of residuals

Residual Check	Spring	Summer	Autumn	Winter
Independence (Durbin-Watson)	2.01 (-0.01)	2.04 (-0.04)	2.11 (-0.06)	2.10 (-0.11)
Homoscedasticity (White)	8.30 (0.50)	8.09 (0.53)	5.46 (0.79)	14.49 (0.11)
Normality (Shapiro-Wilk)	0.93 (0.01)	0.93 (0.01)	0.98 (0.75)	0.94 (0.02)

Table 5. Diagnostic check of residuals for the data transformed using Box-Cox transformation

Residual Check	Spring	Summer	Autumn	Winter
Independence (Durbin-Watson)	2.00 (-0.01)	2.11 (-0.07)	2.11 (-0.06)	1.82 (0.08)
Homoscedasticity (White)	10.36 (0.32)	13.40 (0.15)	5.46 (0.79)	8.56 (0.48)
Normality (Shapiro-Wilk)	0.967 (0.22)	0.953 (0.07)	0.983 (0.75)	0.980 (0.61)

교차확인(cross validation)은 예측오차를 추정하기 위한 표준적인 도구이다(Efron과 Tibshirani, 1993). 회귀모형의 예측능력을 검토하기 위해서 하나의 자료를 제외하고 나머지 자료를 이용하여 회귀모형을 적합 시키고 적합된 회귀모형을 이용하여 제외된 자료를 예측하는 교차확인을 수행하였다. 교차확인에 의한 예측결과를 평균절대오차(mean absolute error, MAE)와 평균제곱오차(mean squared error, MSE) 그리고 기술점수(skill score, SS)를 이용하여 평가하였다.

예측능력은 일반적으로 기술점수에 의해서 제시되며, 이것은 기준예측에 대한 개선정도의 %로서 해석된다. 일반적인 형태로서 예측에 대한 기술점수는 기준예측의 정확도 A_{ref}에 대한 예측정확도 A의 상대적인 크기로 식 (12)와 같이 표현된다(Wilks, 1995).

$$\mathrm{SS}_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\% \tag{12}$$

여기서 A_{perf}는 완전예측(perfect forecasts)에 의해서 달성될 수 있는 정확도 척도의 값이다.

Table 7에 교차확인 방법에 의한 평균제곱오차, 평균절대오차, 그리고 이들에 의한 기술점수를 나타내었다. 여기서 $\mathrm{MSE}_{\mathrm{Clim}}$는 기후학적 평균값을 기준예측으로 했을 때의 평균제곱오차를 나타내며, $\mathrm{MSE}_{\mathrm{Pers}}$는 직전 시간의 값(각 계절별 유입량이므로 봄철이라면 그 이전 해 봄철의 유입량)을 기준예측으로 했을 때를 나타낸다. 이는 평균절대오차에서도 동일하다. 이를 토대로 판단하면 겨울철이 가장 예측력이 높고 여름철이 가장 낮게 나타났다. 그리고 기준예측으로 지속성에 의한 예측을 이용하면 평균을 사용하는 것보다 예측능력 향상 정도가 크게 나타남을 볼 수 있다. 이러한 결과로부터 예측결과가 없을 때 평균 유입량을 사용하는 것이 직전 시간의 유입량을 이용하는 것 보다는 더 안전할 것으로 판단된다.

Table 7. Cross validation skill score of multiple linear regression

Skill	Spring	Summer	Autumn	Winter
MSE	120.27	381.37	109.58	11.38
$\mathrm{MSE}_{\mathrm{Clim}}$	194.83	583.96	205.19	23.53
Skill Score	38.27	34.69	46.60	51.63
$\mathrm{MSE}_{\mathrm{Pers}}$	428.32	1174.77	450.95	36.85
Skill Score	71.92	67.54	75.70	69.12
MAE	8.19	15.33	8.47	2.50
$\mathrm{MAE}_{\mathrm{Clim}}$	10.66	19.44	11.61	3.59
Skill Score	23.15	21.15	27.00	30.28
$\mathrm{MAE}_{\mathrm{Pers}}$	17.65	26.46	16.85	4.02
Skill Score	53.59	42.06	49.71	37.66

Table 6. Fitted regression equations

Season	Fitted Regression Equation
Spring	$\hat{y_i} = \{0.55 \times (-542.5466 + 0.0286x_{i,1} + 0.0452x_{i,2} + 0.0312x_{i,3}) + 1\}^{\frac{1}{0.55}}$
Summer	$\hat{y_i} = \{0.15 \times (-16.5932 + 0.0142x_{i,1} - 1.2371x_{i,2} - 0.8522x_{i,3}) + 1\}^{\frac{1}{0.15}}$
Autumn	$\hat{y_i} = -1082.1592 + 14.1336x_{i,1} - 23.8259x_{i,2} + 0.2382x_{i,3}$
Winter	$\hat{y_i} = \{0.25 \times (-175.7159 + 1.4446x_{i,1} + 0.0174x_{i,2} + 1.5112x_{i,3}) + 1\}^{\frac{1}{0.25}}$

예측결과의 평가를 3개의 범주를 갖는 자료로 변환하여 수행해 보았다. 예측결과를 범주형 예측으로 변환하기 위한 기준구간은 Table 8과 같다. 기준구간을 설정한 방법은 자료를 크기순으로 정렬한 다음 순위를 전체 자료수로 나눈 값으로 확률을 부여하고, 카테고리의 확률이 1/3이 되는 지점의 경계값을 계산한 것이다. Table 9는 Table 8을 이용하여 범주형으로 변환된 예측결과에 대한 분할표를 나타낸 것이다. 분할표를 이용하여 회귀분석에 의한 계절별 유입량 예측 결과를 카테고리 예측으로 변환하여 평가를 수행하였다. 가장 단순한 카테고리 예측은 [있음, 없음]의 두 가지 형태로 예측결과를 제시하는 것이다. 따라서 어떤 기준값을 넘는 값이 발생할 것인지에만 관심이 있으며, 그 이외에 카테고리의 확률이나 예측값은 제공되지 않는 단순한 형태의 예측이라고 할 수 있다. 예측결과가 카테고리 예측으로 주어지는 경우에는 Heidke 기술점수(Heidke skill score, HSS)로 예측능력을 평가할 수 있다. HSS는 기본적인 정확도 척도로서 적중률에 기초하고 있으며, 다음 식 (13)과 같다(Wilks, 1995).

Table 8. Reference values for categorization (unit: m^3/s)

Category	Spring	Summer	Autumn	Winter
Below Normal	$y < 15.88$	$y < 46.43$	$y < 15.48$	$y < 3.47$
Normal	$15.88 \leqq y < 27.40$	$46.43 \leqq y < 67.75$	$15.48 \leqq y < 28.52$	$3.47 \leqq y < 6.63$
Above Normal	$y \geqq 27.40$	$y \geqq 67.75$	$y \geqq 28.52$	$y \geqq 6.63$

Table 9. Contingency table of cross validation: multiple linear regression

Spring				
Forecast \ Observation	Below Normal	Normal	Above Normal	Total
Below Normal	6	3	3	12
Normal	9	8	4	21
Above Normal	0	4	8	12
Total	15	15	15	45
Summer				
Forecast \ Observation	Below Normal	Normal	Above Normal	Total
Below Normal	11	3	3	17
Normal	4	7	7	18
Above Normal	0	5	5	10
Total	15	15	15	45
Autumn				
Forecast \ Observation	Below Normal	Normal	Above Normal	Total
Below Normal	5	5	1	11
Normal	8	8	2	18
Above Normal	2	2	12	16
Total	15	15	15	45
Winter				
Forecast \ Observation	Below Normal	Normal	Above Normal	Total
Below Normal	10	4	1	15
Normal	5	6	4	15
Above Normal	0	5	10	15
Total	15	15	15	45

$$\mathrm{HSS}=\frac{\sum_{i=1}^{I}p(y_i,o_i)-\sum_{i=1}^{I}p(y_i)p(o_i)}{1-\sum_{i=1}^{I}p(y_i)p(o_i)} \quad (13)$$

식 (8)에서 y_i는 예측값, o_i는 관측값을 의미하며, HSS는 완전예측의 경우 1, 기준예측과 같으면 0, 기준예측보다 나쁘면 음의 점수를 갖는다. 예측의 평가를 위해서 Heidke 점수를 계산하면 다음 Table 10과 같다. 기준예측에 비해서 23%~37% 향상된 것으로 나타났으며, 예측의 향상정도는 겨울, 가을, 여름, 그리고 봄의 순서로 나타났다. 다른 계절에 비해서 상대적으로 봄철의 유입량을 예측하는 것이 쉽지 않은 것으로 해석할 수 있다.

Table 10. Heidke skill score of the categorical forecast using multiple linear regression

Skill Score	Spring	Summer	Autumn	Winter
Heidke Skill Score	0.23	0.27	0.33	0.37

4. 결 론

본 연구에서는 해수면온도와 북반구 500 hPa 지위고도 자료로부터 예측인자를 선정하여 다중선형회귀분석으로 안동댐 계절별 유입량 예측을 수행하였다. 회귀모형의 예측능력을 검토하기 위해서 교차확인을 수행하여 교차확인에 의한 예측결과를 평균절대오차와 평균제곱오차 그리고 기술점수를 이용하여 평가하였다. 또한 회귀분석 예측결과를 카테고리 예측으로 변환하여 분할표를 작성하고 이를 Heidke 기술점수를 토대로 평가하였다.

예측인자를 선정할 때 교차상관계수와 상호정보량을 보완적으로 사용하여 이들 자료로부터 1차적으로 예측인자를 식별하였다. 그리고 1차적으로 선택된 예측인자 집합으로부터 부분 교차상관계수와 부분 상호정보량을 이용하여 최종적인 예측인자를 식별하였다. 안동댐 계절별 유입량과 남방진동지수와는 유의한 상관관계가 발견되지 않았다. 해수면온도와 지위고도자료는 안동댐 계절별 유입량과 유의한 관계를 갖는 지점이 많이 발견되었다. 따라서 해수면온도와 지위고도자료는 안동댐 계절별 유입량 예측에 중요한 인자라고 할 수 있다. 일차적으로 교차상관계수와 상호정보량을 통하여 선택된 예측인자들 중에서 부분 교차상관계수와 부분상호정보량을 이용하여 예측인자를 추출한 결과 각 계절별로 3개씩 선택되었다. 예측인자의 선정과정에 (부분) 교차상관계수와 (부분) 상호정보량을 상호보완적으로 사용하였는데 이는 한 가지 방법에 의해서 잘못된 선택을 할 수 있는 가능성을 줄일 수 있는 방법이라 생각된다. 그리고 이들의 유의수준을 붓스트랩에 의한 방법으로 설정한 부분도 통계패키지와 비교한 결과 타당성이 있다고 판단된다.

식별된 예측인자를 이용하여 다중선형회귀모형을 적용하여 계절별로 예측을 수행하고 평가하였다. 교차확인의 예측결과를 평균절대오차와 평균제곱오차로 평가한 결과 기준예측에 비해서 향상된 것을 볼 수 있었다. 회귀분석결과를 카테고리로 나누어 분할표를 작성하고 이를 Heidke 기술점수를 토대로 평가하였으며, 그 결과 기후학적 예측보다 향상된 결과를 나타내었다. Heidke 기술점수를 통해서 카테고리 예측은 기준예측에 비해서 23%~37% 향상된 것으로 나타났으며, 예측의 향상정도는 겨울, 가을, 여름, 그리고 봄의 순서로 나타났는데 다른 계절에 비해서 상대적으로 봄철의 유입량을 예측하는 것이 쉽지 않은 것으로 해석할 수 있다.

다양한 예측기간에 대한 실제 예측가능성을 파악하고 예측결과의 정확도가 어느 정도 이상 되었을 경우 불확실성을 포함한 결과가 댐 운영이나 물 공급 및 수요관리 등 현실적인 물관리에 도움이 될 수 있는지에 대한 분석은 추후 더욱 연구가 필요한 부분이라고 판단된다.

감사의 글

본 연구는 국토교통부 물관리연구사업의 연구비지원(11기술혁신C06)에 의해 수행되었습니다.

참 고 문 헌

Ahn, J. B., Park, J. Y., 2000, A Study on a Statistical Long-term Prediction Model Using Global Sea-Surface

Temperature Anomalies, Asia-Pacific Journal of Atmospheric Sciences, 36(2), pp. 179-188.

Awadallah, A.G., Rousselle, J., 2000, Improving Forecasts of Nile Flood Using SST Inputs in TFN Model, Journal of Hydrologic Engineering, 5(4), pp. 371-379.

Berri, G.J., Flamenco, E.A., 1999, Seasonal Volume Forecast of the Diamante River, Argentina, Based on El Nino Observations and Predictions, Water Resources Research, 35(12), pp. 3803-3810.

Chiew, F.H.S., Piechota, T.C., Dracup, J.A., McMaho, T.A., 1998, El Niño/Southern Oscillation and Australian Rainfall, Streamflow and Drought: Links and Potential for Forecasting, Journal of Hydrology, 204, pp. 138-149.

Climate Prediction Center: http://www.cpc.ncep.noaa.gov/

Cover, T.M., Thomas, J.A., 1991, Elements of Information Theory, John Wiley & Sons, Inc.

Draper, N.R., Smith, H., 1998, Applied Regression Analysis, 3rd Edition, John Wiley & Sons.

Efron, B., Tibshirani, R.J., 1993, An Introduction to the Bootstrap, Chapman & Hall.

Hipel, K.W., McLeod, A.I., 1994, Time Series Modelling of Water Resources and Environmental Systems, Elservier.

Hocking, R.R., 1996, Methods and Applications of Linear Models: Regression and the Analysis of Variance, John Wiley & Sons.

Liu, Z., Valdes, J.B., Entekhabi, D., 1998, Merging and Error Analysis of Regional Hydrometeorologic Anomaly Forecast Conditioned on Climatic Precursors, Water Resources Research, 34(8), pp. 1959-1969.

Piechota, T.C., Chiew, F.H.S., Dracup, J.A., McMahon, T.A., 1998, Seasonal Streamflow Forecasting in Eastern Australia and the El Nino-Southern Oscillation, Water Resources Research, 34(11), pp. 3035-3044.

Piechota, T.C., Chiew, F.H.S., Dracup, J.A., McMahon, T.A., 2001, Development of Exceedance Probability Streamflow Forecast, Journal of Hydrologic Engineering, 6(1), pp. 20-28.

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, pp. 232-239.

Sharma, A. Luk, K.C., Cordery, I., Lall, U., 2000, Seasonal to Interannual Rainfall Probabilistic Forecasts for Improved Water Supply Management: Part 2 - Predictor Identification of Quarterly Rainfall Using Ocean-Atmosphere Information, Journal of Hydrology, 239, pp. 232-239.

Simpson, H.J., Cane, M.A., Herczeg, A.L., Zebiak, S.E., Simpson, J.H., 1993, Annual River Discharge in Southeastern Australia Related to El Niño-Southern Oscillation Forecasts of Sea Surface Temperatures, Water Resources Research, 34(11), pp. 3035-3044.

Simpson, H.J., Colodner, D.C., 1999, Arizona Precipitation Response to the Southern Oscillation: A Potential Water Management Tool, Water Resources Research, 35(12), pp. 3761-3769.

Uvo, C.B., Graham, N.E., 1998, Seasonal Runoff Forecast for Northern South America: A Statistical Model, Water Resources Research, 34(12), pp. 3515-3524.

Wei, W.W.S., 1990, Time Series Analysis: Univariate and Multivariate Methods, Addison-Wesley.

Wilks, D.S., 1995, Statistical Methods in the Atmospheric Sciences: An Introduction, Academic Press.