PE-98972

연안 Air-Sea 경계에서의 다변수 상호작용 시계열 모델링

Time-series modeling of the multi-variate interactions in a coastal air-sea interface

2014. 6. 30.

한국해양과학기술원

제 출 문

한국해양과학기술원장 귀하

본 보고서를 “연안 Air-Sea 경계에서의 다변수 상호작용 시계열 모델링” 과제의 (연차,최종)보고서로 제출합니다.

2014. 6. 30.

총괄연구책임자 : 조 홍 연

참 여 연 구 원 : 김 성
참 여 연 구 원 : 김 경 옥
참 여 연 구 원 : 조 범 준
참 여 연 구 원 : 방 윤 경

보고서 초록

<table>
<tr><td>과제고유
번호</td><td>PE98972</td><td>해당단계
연구기간</td><td></td><td>단계 구분</td><td>최종보고서</td></tr>
<tr><td rowspan="2">연구사업명</td><td>중사업명</td><td colspan="4"></td></tr>
<tr><td>세부사업명</td><td colspan="4"></td></tr>
<tr><td rowspan="2">연구과제명</td><td>대과제명</td><td colspan="4">창의연구사업 (Lab. 창의과제)</td></tr>
<tr><td>세부과제명</td><td colspan="4">연안 Air-Sea 경계에서의 다변수 상호작용 시계열 모델링</td></tr>
<tr><td rowspan="2">연구책임자</td><td rowspan="2">조 홍 연</td><td>해당단계
참여연구원수</td><td>총 : 명
내부: 명
외부: 명</td><td>해당단계
연구비</td><td>정부: 천원
기업: 천원
계 : 천원</td></tr>
<tr><td>총연구기간
참여연구원수</td><td>총 : 5 명
내부: 5 명
외부: 명</td><td>총
연구비</td><td>정부: 50,000 천원
기업: 천원
계 : 50,000 천원</td></tr>
<tr><td>연구기관명
및
소속부서명</td><td colspan="2">한국해양과학기술원
해양환경보전연구부</td><td>참여기업명</td><td colspan="2"></td></tr>
<tr><td>국제공동연구</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>위탁연구</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="4">요약 (연구결과를 중심으로 개조식 500자 이내)</td><td>보고서
면수</td><td></td></tr>
<tr><td colspan="6">해양에서 부이 등을 이용하여 관측되는 자료에는 이상자료 및 결측이 빈번하게 발생.
관측자료를 이용하여 모델 신뢰수준 및 예측성능 평가 등이 수행되기 때문에 완전한 자료가 필요
- 연안 모니터링 자료의 잔차성분을 이용하여 이상자료를 탐지하는 기술 개발-적용, 잔차성분 도출과정은 비모수적 방법에 해당하는 최적의 Kernel regression smoothing 기법을 이용하여 연속적인 어림성분을 도출하고, 어림성분과 관측자료의 차이를 잔차성분(나머지 성분, detailed components)으로 간주하는 기법을 개발-보완-적용.
- 연안 모니터링 자료의 잔차성분을 이용하여 결측 구간 자료를 보충하는 기법 개발-적용
- 관련 기법의 논문 게재(Hong-Yeon Cho, Jihee Oh, Kyeongok Kim, and Jae-Seol, Shim, 2013. 4. Outlier detection and missing data filling methods for coastal water temperature data, Journal of Coastal Research, Special Issue, No. 65, pp.1898-1903.)
- 처리되어 완전한 자료로 구성된 수온 자료를 이용하여 장기(1961-2100: 140년) 수온 자료를 구성
- 관련 연구 성과의 논문 투고(심사진행 - Khil-Ha Lee, Hong-Yeon Cho, Projection of Climate-Induced Future Water Temperature for Aquatic Environment, J. of Environmental Engineering, ASCE)</td></tr>
<tr><td rowspan="2">색인어
(각 5개 이상)</td><td>한 글</td><td colspan="4">시계열모델링, 대기-해양경계, 이상자료, 결측보충, 완전한 자료</td></tr>
<tr><td>영 어</td><td colspan="4">time-series modeling, air-sea interface, outlier, missing imputation, complete data</td></tr>
</table>

요 약 문

Ⅰ. 제 목

연안 Air-Sea 경계에서의 다변수 상호작용 시계열 모델링

Ⅱ. 연구개발의 목적 및 필요성

- 연구개발의 목적

○ 연안 Air-Sea 경계에서의 에너지 수지분석 기술수준 향상

○ 연안에서의 에너지 수지분석에 영향을 미치는 인자의 시계열 자료 구축

○ Air-Sea 경계에서 열수지 분석에 영향을 미치는 주요한 모든 영향인자를 예측하는 모형의 개발 및 성능평가(추정 오차 및 신뢰구간)

- 연구개발의 필요성

가. 기술적 측면

○ Air-Sea 경계에서는 기온과 수온, 태양복사 및 지구복사에너지, 현열, 잠열이 서로 복잡하게 상호 영향을 미치고 있으나, 모든 자료가 가용한 상태가 아니며 주로 태양복사에너지자료와 기본적인 기상자료(풍속, 습도 등)만을 관측하여 다른 모든 인자를 일방향으로 추정하고 있기 때문에 정확한 현상재현이 곤란한 상태이기 때문에 기후변화에 따른 연안 환경변화 예측이 곤란한 상태임.

○ 대기-해수면 경계에서의 열/에너지 수지(heat/energy budget)는 매우 다양한 기상변수와 수온이 관련되어 있으며 상호작용(interaction)을 하고 있으나, 태양복사 등이 일방향으로 영향을 미치는 방식으로 모의하는 경우가 대부분이며, 대기에서나 해양에서 경계면으로 매우 중요하기 때문에 정확한 경계면 정보구성이 필요.

○ 한편, 정확한 Air-Sea 경계정보 구성을 위하여 보조적인 관측이 수행될 수 있으나 모든 필요한 정보를 관측을 통하여 확보하는 방법에는 실질적인 한계가 있기 때문에 기존의 가용한 자료(국립수산과학원 수온 정점관측자료 및 어장정보시스템관측자료, 국립해양조사원 기온-수온 관측자료 및 부이 관측자료, 기상청 부이-등표 자료 등)를 최대한 활용하여 통계 분석을 수행하여 완전한 자료(complete data)를 구성하여 정보를 추출하는 방법이 필요.

* 연안 Air-Sea 경계에서 에너지 수지분석에 영향을 미치는 인자의 자료 상황.

1. 수면 도달 태양복사자료 - 인접한 육지에서의 기상자료가 가용한 상황.

2. 수면 도달 장파복사 자료 - 일조시간 및 습도의 함수로 표현되는 공식을 사용하여 추정.

3. 현열 및 잠열 - 기온과 수온의 차이 및 포화증기압(습도의 함수), 풍속의 함수로 표현되는 공식을 이용하여 추정. 기온과 수온자료를 이용하여 에너지를 추정하고, 그 에너지 수지가 다시 기온 및 수온에 영향을 미치기 때문에 상호작용을 고려하여야 함.

4. 기온 및 수온 - 해양에서는 장파복사자료를 이용한 해수면온도 자료가 가용하지만, 연안에서는 공간적인 해수면온도 분포자료가 미흡한 상황. 정점 관측 자료로 제한되어 있는 상황. 주요 연안에 대하여 수치모델링 기법 등을 통한 공간분포 추정 작업도 필요. 반면 기온도 연안에서 직접 관측한 자료도 부족한 상황이나 인근 기상자료가 장기간 가용한 상황.

나. 경제, 산업적 측면

○ 기후변화에 따른 환경변화는 연안 환경에 직접적인 영향을 미칠 것으로 판단되며, 연안 환경변화는 생태환경변화 및 수산자원에 영향을 미치기 때문에 해양 환경산업 및 수산분야에서 매우 필수적인 주제로 정확한 연구가 필요.

○ 기후산업이 부각되는 상황에서 해양산업도 연안기상 및 환경변화 정도가 산업영역에도 크게 영향을 미칠 것으로 예상된다. 이러한 측면에서 보다 정확하고 실용적인 연안 환경정보, 그 중에서도 기온 및 수온정보 제공은 매우 기본적이고 필수적인 자료에 해당하기 때문에 필요.

다. 사회, 문화적 측면

○ 기후변화에 따른 연안 환경변화는 최근 부각되는 연안을 중심으로 부각되는 다양한 휴양·레져 활동을 제한하는 요소이기 때문에 연안 관광 등의 장기적인 정책 결정을 위한 기초자료로서 필요함.

Ⅲ. 연구개발의 내용 및 범위

○ Air & Sea 경계에서 에너지 수지분석에 영향을 미치는 모든 인자의 상호작용을 고려한 시계열 모델링

○ 다변수 시계열 모델링 기법을 이용하여 경계면 결측자료(missing data) 및 이상자료(outlier)를 제거-보충하여 신뢰할 수 있는 완전한(complete) 자료 구성 기법 개발-적용

○ 우리나라 주요 연안에서 가용한 자료를 기반으로 한 최소 30년 이상의 Air & Sea 경계 영향인자 시계열 자료 재구축(re-construction)

Ⅳ. 연구개발결과

- 연안 모니터링 자료의 잔차성분을 이용하여 이상자료를 탐지하는기술 개발-적용

- 연안 모니터링 자료의 잔차성분을 이용하여 결측 구간 자료를 보충하는 기법 개발-적용

- 관련 기법의 논문 게재(Hong-Yeon Cho, Jihee Oh, Kyeongok Kim, and Jae-Seol, Shim, 2013. 4. Outlier detection and missing data filling methods for coastal water temperature data, *Journal of Coastal Research*, Special Issue, No. 65, pp.1898-1903.)

- 처리되어 완전한 자료로 구성된 수온 자료를 이용하여 장기(1961-2100: 140년) 수온자료를 구성, 구성지점은 시화호, 마산만, 낙동강 하구지점.

- 관련 연구 성과의 논문 투고(심사진행 - Khil-Ha Lee, Hong-Yeon Cho, Projection of Climate-Induced Future Water Temperature for Aquatic Environment, J. of Environmental Engineering, ASCE)

Ⅴ. 연구개발결과의 활용계획

○ 기후변화에 따른 연안 환경변화 예측 및 생태변화 예측분야에 활용될 수 있기 때문에 해양환경변화와의 접목을 통하여 국가 연구개발사업으로 발전하는 것이 가능.

○ 연안 수온은 그 중요도에 비하여 자료구축 수준이 상대적으로 기상인자에 비하여 매우 미약한 수준이기 때문에 본 연구에서 개발된 기법을 이용하여 자료를 구축하고, 구축된 자료를 이용하여 장래 수온을 포함한 환경인자의 변화 예측 등을 통하여 장기적인 연안 환경-생태계-수산자원 관리 분야에 활용할 수 있기 때문에 우리 기술원 주도의 연안 생태환경변화를 주도할 수 있음.

S U M M A R Y

Ⅰ. Title

Time-series modeling of the multi-variate interactions in a coastal air-sea interface

Ⅱ. The goal and necessity of research

1. The goal of research

○ Improvement of the heat-budget analysis level in the air-sea interface

○ Time-series data reconstruction of the influence factors on the heat (energy) budget mechanism in the coastal seas.

○ Development and performance test (estimation errors and confidence interval) of the model predicting the major parameters closely related to the heat balance in the air-sea interface.

2. The necessity of research

(1) Technical aspect

○ There are close and complex interactions between air and water temperatures, solar (short-wave) radiation, long-wave radiation, sensible-heat, and evaporative heats in the air-sea interface. Only few main parameters, such as air and water temperatures and solar radiation, are observed. The other parameters should be estimated using the available empirical formula and relationships. It makes the predicted heat balance uncertain in a certain level.

○ There are basic interactions between the related parameters in the heat/energy transfer mechanism in the air-sea interface. It is, however, considered as the one-way mechanism.

○ It can be carried out to observe the additional parameters related to the heat balance for the more accurate air-sea interface information of the parameters. However, it is practically limited. The method using the available data as possible as we collect is the most suitable and effective method. The available data sets are the KHOA monitoring data, NFRDI monitoring data, and KMA buoy monitoring data. The data set should be checked for the quality assurance, outlier detection and missing

imputation.

(2) Economic and industrial aspect

○ Coastal environmental change are essential issues because it is directly effected to the ecological change and fishery management.

○ Providing of the more accurate and practical coastal information, especially air and water temperatures, can be widely used to the marine leisure industry.

(3) Social and cultural aspect

○ It is highly required as the basic data (data-infra.) for the long-term policy-making of the coastal tourism, leisure-sports, and so on.

Ⅲ. Contents and scopes of research

○ Time-series modeling considering the every parameter's interaction related to the heat budget analysis in the air-sea interface.

○ Development and application of the reconstruction method on the reliable complete data through the outlier detection/remova and missing imputation (filling-in) using the multi-variate time-series modeling.

○ Long-term (140 years) time-series water temperature data reconstruction using the air-water temperatures relationship model and available monitoring data in the coastal seas.

Ⅳ. Research outcome

- Development and application of the outlier detection method using the residual (detailed) components of the coastal monitoring data

- Development and application of the missing imputation (filling-in) using the statistical information of the residual components of the coastal monitoring data.

- Published article : Hong-Yeon Cho, Jihee Oh, Kyeongok Kim, and Jae-Seol, Shim, 2013. 4. Outlier detection and missing data filling methods for coastal water temperature data, *Journal of Coastal Research*, Special Issue, No. 65, pp.1898-1903.)

- Reconstruction of the 140 years (1961-2100) complete coastal air and water temperatures data set.

- Submitted articles : Khil-Ha Lee, Hong-Yeon Cho, Projection of Climate-Induced Future Water Temperature for Aquatic Environment, J. of Environmental Engineering, ASCE.

Ⅴ. Plans of applying research outcome

○ It can be used in the coastal environmental and/or ecological fields due to the climate change. It is possible to expand the national projects because the complete

coastal data set can be widely used to the coastal research fields.

○ The KIOST become a leading group of the coastal environmental/ecological research fields as the data-hub center in the era of the big-data.

Keywords :
<국문> 시계열모델링, 대기-해양경계, 이상자료, 결측구간의 자료보충, 완전한 자료
<영문> time-series modeling, air-sea interface, outlier, missing imputation, complete data

C O N T E N T S

목 차

제 1 장 서론

1. 연구개발의 목적

○ 연안 Air-Sea 경계에서의 에너지 수지분석 기술수준 향상

○ 연안에서의 에너지 수지분석에 영향을 미치는 인자의 시계열 자료 구축

○ Air-Sea 경계에서 열수지 분석에 영향을 미치는 주요한 모든 영향인자를 예측하는 모형의 개발 및 성능평가(추정 오차 및 신뢰구간)

2. 연구개발의 필요성

가. 기술적 측면

○ Air-Sea 경계에서는 기온과 수온, 태양복사 및 지구복사에너지, 현열, 잠열이 서로 복잡하게 상호 영향을 미치고 있으나, 모든 자료가 가용한 상태가 아니며 주로 태양복사 에너지자료와 기본적인 기상자료(풍속, 습도 등)만을 관측하여 다른 모든 인자를 일방향으로 추정하고 있기 때문에 정확한 현상재현이 곤란한 상태이기 때문에 기후변화에 따른 연안 환경변화 예측이 곤란한 상태임.

○ 대기-해수면 경계에서의 열/에너지 수지(heat/energy budget)는 매우 다양한 기상변수와 수온이 관련되어 있으며 상호작용(interaction)을 하고 있으나, 태양복사 등이 일방향으로 영향을 미치는 방식으로 모의하는 경우가 대부분이며, 대기에서나 해양에서 경계면으로 매우 중요하기 때문에 정확한 경계면 정보구성이 필요.

○ 한편, 정확한 Air-Sea 경계정보 구성을 위하여 보조적인 관측이 수행될 수 있으나 모든 필요한 정보를 관측을 통하여 확보하는 방법에는 실질적인 한계가 있기 때문에 기존의 가용한 자료(국립수산과학원 수온 정점관측자료 및 어장정보시스템관측자료, 국립해양조사원 기온-수온 관측자료 및 부이 관측자료, 기상청 부이-등표 자료 등)를 최대한 활용하여 통계 분석을 수행하여 완전한 자료(complete data)를 구성하여 정보를 추출하는 방법이 필요.

* 연안 Air-Sea 경계에서 에너지 수지분석에 영향을 미치는 인자의 자료 상황.

1. 수면 도달 태양복사자료 - 인접한 육지에서의 기상자료가 가용한 상황.

2. 수면 도달 장파복사 자료 - 일조시간 및 습도의 함수로 표현되는 공식을 사용하여 추정.

3. 현열 및 잠열 - 기온과 수온의 차이 및 포화증기압(습도의 함수), 풍속의 함수로 표현되는 공식을 이용하여 추정. 기온과 수온자료를 이용하여 에너지를 추정하고, 그 에너지 수지가 다시 기온 및 수온에 영향을 미치기 때문에 상호작용을 고려하여야 함.

4. 기온 및 수온 - 해양에서는 장파복사자료를 이용한 해수면온도 자료가 가용하지만,

연안에서는 공간적인 해수면온도 분포자료가 미흡한 상황. 정점 관측 자료로 제한되어 있는 상황. 주요 연안에 대하여 수치모델링 기법 등을 통한 공간분포 추정 작업도 필요. 반면 기온도 연안에서 직접 관측한 자료도 부족한 상황이나 인근 기상자료가 장기간 가용한 상황.

나. 경제, 산업적 측면

○ 기후변화에 따른 환경변화는 연안 환경에 직접적인 영향을 미칠 것으로 판단되며, 연안 환경변화는 생태환경변화 및 수산자원에 영향을 미치기 때문에 해양 환경산업 및 수산분야에서 매우 필수적인 주제로 정확한 연구가 필요.

○ 기후산업이 부각되는 상황에서 해양산업도 연안기상 및 환경변화 정도가 산업영역에도 크게 영향을 미칠 것으로 예상된다. 이러한 측면에서 보다 정확하고 실용적인 연안환경정보, 그 중에서도 기온 및 수온정보 제공은 매우 기본적이고 필수적인 자료에 해당하기 때문에 필요.

다. 사회, 문화적 측면

○ 기후변화에 따른 연안 환경변화는 최근 부각되는 연안을 중심으로 부각되는 다양한 휴양·레져 활동을 제한하는 요소이기 때문에 연안 관광 등의 장기적인 정책 결정을 위한 기초자료로서 필요함.

3. 연구개발의 내용 및 범위

○ Air & Sea 경계에서 에너지 수지분석에 영향을 미치는 모든 인자의 상호작용을 고려한 시계열 모델링

○ 다변수 시계열 모델링 기법을 이용하여 경계면 결측자료(missing data) 및 이상자료(outlier)를 제거-보충하여 신뢰할 수 있는 완전한(complete) 자료 구성 기법 개발-적용

○ 우리나라 주요 연안에서 가용한 자료를 기반으로 한 최소 30년 이상의 Air & Sea 경계 영향인자 시계열 자료 재구축(re-construction)

제 2 장 국내외 기술개발 현황

가. 국외

○ 국외에서는 해수 수온 예측모형의 개발 측면에서 경계면에서의 열수지 분석을 통한 일방향(영향인자간의 상호작용을 고려하지 않는 모형) 예측모델 개발이 수행. 장파복사자료를 이용한 해수면 온도 관측자료 분석이 대부분을 차지하고 있음.

○ 또한, 해수면 표피층(skin layer) 등을 포함하는 미시적인 관점에서의 열전달 등의 연구가 수행되고 있음.

나. 국내

○ 국내에서는 태양복사량 자료를 이용한 일방향 해수 수온 예측모델 개발로 제한되어 있음. 그러나 수온 예측에 중요한 장파복사에 대한 정보는 지역적인 변화가 크게 나타나고 있음에도 불구하고 외국의 경험공식에 의존하고 있은 상태이며, 해수 유동의 관점에서 혼합층 중심으로 수행되었음.

○ 또한 기후변화에 따른 해양환경의 변화가 예상되는 상황이지만, 육상 기상자료에 비하여 해양환경자료는 빈약하기 때문에 대부분의 자료를 추정하여 해양 수온 환경변화 등을 예측하는 연구가 대부분을 차지하고 있다.

제 3 장 연구개발수행 내용 및 결과

1. 연구수행 추진체계

가. 제1단계

○ 우리나라 연안의 가용한 모든 기온 및 수온자료를 중심으로 Air-Sea 경계에서의 에너지수지에 영향을 미치는 자료를 수집-정리

○ 수집된 자료를 이용한 자료의 전처리 기법 적용을 통한 품질 평가

○ 특별한 추진전략 및 체계보다는 가용한 자료를 이용하여 현재 가용한 통계기법을 적용하여 기후변화에 따른 연안 환경변화, 특히 수온 등의 변화를 1.0 ℃ 이내로 정확하게 예측할 수 있는 시계열 모델개발에 있음.

나. 제2단계

○ 시간적·공간적 통계적 모델링 기법을 이용한 다변량 시계열 모델링 수행

○ 특별한 추진전략 및 체계보다는 가용한 자료를 이용하여 현재 가용한 통계기법을 적용하여 기후변화에 따른 연안 환경변화, 특히 수온 등의 변화를 1.0 ℃ 이내로 정확하게 예측할 수 있는 시계열 모델개발에 있음.

2. 연구수행 결과

- 연안 모니터링 자료의 잔차성분을 이용하여 이상자료를 탐지하는 기술 개발-적용
- 연안 모니터링 자료의 잔차성분을 이용하여 결측 구간 자료를 보충하는 기법 개발-적용

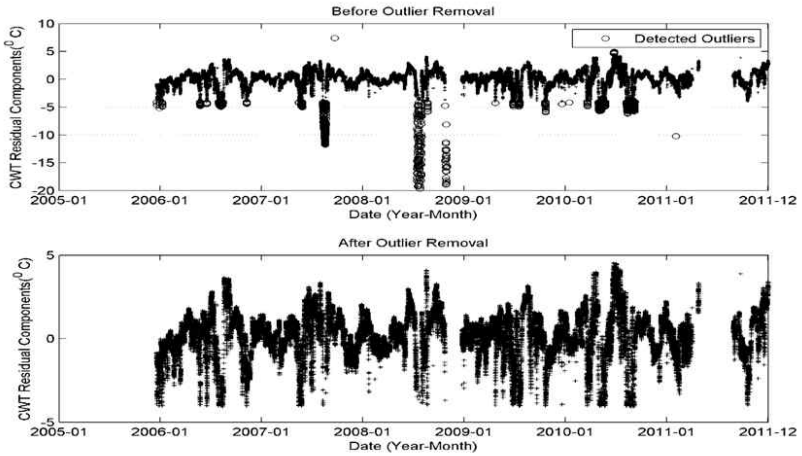


그림 1. 관측 수온자료의 잔차 성분과 감지된 이상자료

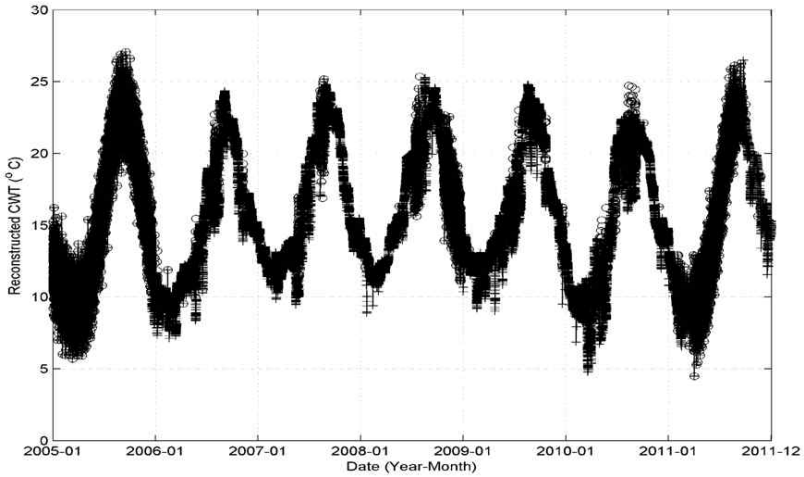


그림 2. 이상자료를 제거하고 결측구간을 보충한 완전한 수온 자료

- 관련 기법의 논문 게재(Hong-Yeon Cho, Jihee Oh, Kyeongok Kim, and Jae-Seol, Shim, 2013. 4. Outlier detection and missing data filling methods for coastal water temperature data, *Journal of Coastal Research*, Special Issue, No. 65, pp.1898-1903.)
- 처리되어 완전한 자료로 구성된 수온 자료를 이용하여 장기(1961-2100: 140년) 수온 자료를 구성

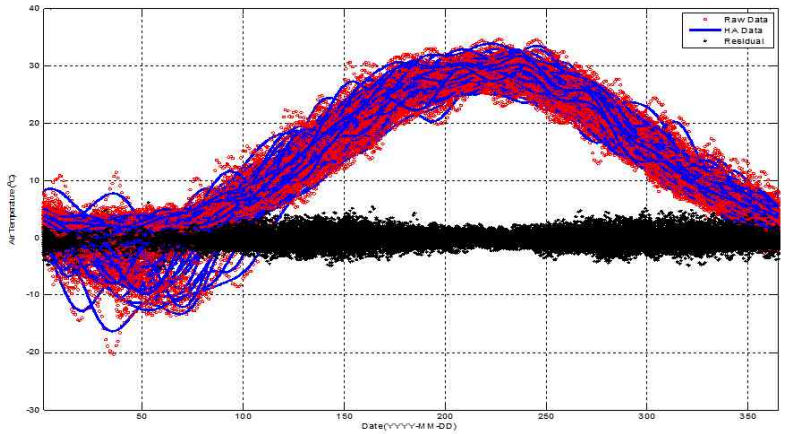


그림 3. 관측자료의 어림성분과 잔차성분 도시 (잔차 성분의 변동크기가 시간에 따라 차이를 보이고 있음. 동계, 하계에는 작고, 춘계-추계에는 큼).

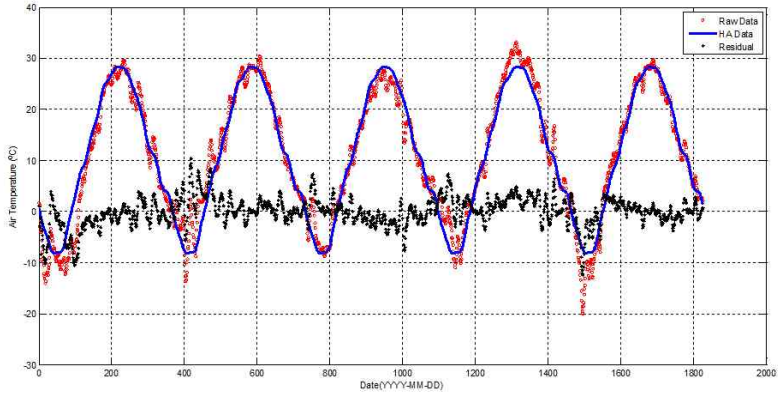


그림 4. 관측 수온자료의 조화분석을 이용한 어림성분과 잔차성분 도출-도시

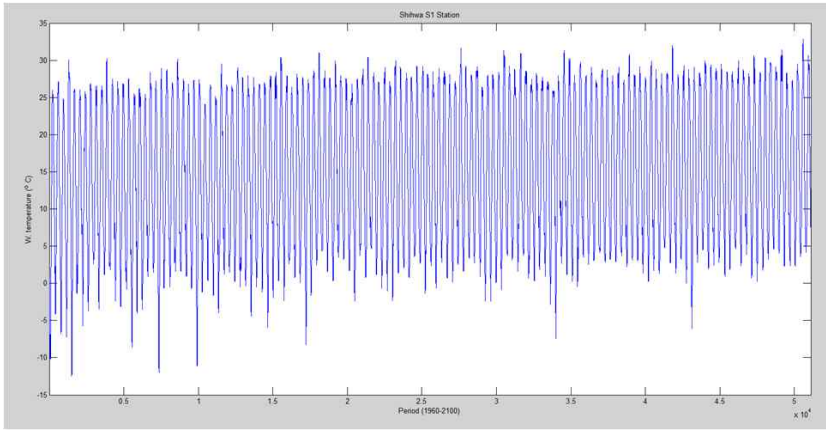


그림 5. 시화호의 140년 동안의 완전한 재구성 자료(S1 지점)

- 관련 연구 성과의 논문 투고(심사진행 - Khil-Ha Lee, Hong-Yeon Cho, Projection of Climate-Induced Future Water Temperature for Aquatic Environment, J. of Environmental Engineering, ASCE)

제 4 장 연구개발목표 달성도 및 대외기여도

1. 대외기여도

가. 기술적 측면

○ 연안 Air-Sea 경계에서의 에너지수지분석 수준 향상

○ 연안에서의 에너지 수지분석을 위한 장기간 영향인자 정보제공(공유)를 통한 연안 환경변화 예측수준 향상 및 생태환경변화 예측을 위한 기초정보 제공가능.

○ 연안 환경자료의 이상자료 감지(outlier detection) 및 결측구간의 자료 보충(missing data imputation) 기법 적용을 통한 완전한 자료 구축으로 다양한 통계분석 연구를 지원.

나. 경제・산업적 측면

○ 연안에서의 환경산업이 부각되는 상황에서 보다 신뢰할 수 있는 환경정보를 제공하여 연안 환경산업 증진에 기여.

제 5 장 연구개발결과의 활용계획

○ 기후변화에 따른 연안 환경변화 예측 및 생태변화 예측분야에 활용될 수 있기 때문에 해양환경변화와의 접목을 통하여 국가 연구개발사업으로 발전하는 것이 가능.

○ 연안 수온은 그 중요도에 비하여 자료구축 수준이 상대적으로 기상인자에 비하여 매우 미약한 수준이기 때문에 본 연구에서 개발된 기법을 이용하여 자료를 구축하고, 구축된 자료를 이용하여 장래 수온을 포함한 관란 환경인자의 변화 예측 등을 통하여 장기적인 연안 환경-생태계-수산자원 관리분야에 활용할 수 있기 때문에 우리 기술원 주도의 연안 생태환경변화를 주도할 수 있음.

제 6 장 참고문헌

Hong-Yeon Cho, Jihee Oh, Kyeongok Kim, and Jae-Seol, Shim, 2013. 4. Outlier detection and missing data filling methods for coastal water temperature data, *Journal of Coastal Research*, Special Issue, No. 65, pp.1898-1903.

정신택, **조홍연**, 고동희 오남선, 손경표, 2013. 2. 한반도 연안 수온자료의 확률분포함수 추정, 한국해안·해양공학회논문집, 제25권, 제1호, 11-19.

Lee, K.H. and **Cho, H.Y.**, 2011. 10. Estimation of the Reference Evapotranspiration using Daily Sunshine Hour, 환경영향평가, 제20권, 제5호, pp.627-640.

이길하, **조홍연**, 2011. 10. 수질영향평가의 신뢰수준 향상을 위한 기상자료의 검정, 환경영향평가, 제20권, 제5호, pp.601-613.

Hongyeon, Cho and Khil-ha, Lee. Development of an air-water temperature relationship model to predict climate-induced future water temperature in estuaries, *J. of Environmental Engineering -ASCE* (Accepted). (DOI : 10.1061/(ASCE)EE.1943-7870.0000499).

Khil-Ha Lee, Hong-Yeon Cho, 2014. (Under Review). Projection of Climate-Induced Future Water Temperature for Aquatic Environment, J. of Environmental Engineering, ASCE.

부록

- 연구과제 수행과 관련하여 출판-심사가 진행되고 있는 논문

부록 1.

조홍연, 오지희, 2012. 5. 어림과 나머지 성분을 이용한 연안 수온자료의 이상자료 감지, 기술보고, 한국해양환경공학회지, 15(2), 156-162.

부록 2.

Hong-Yeon, Cho and Khil-ha, Lee, 2012. 5. Development of an air-water temperature relationship model to predict climate-induced future water temperature in estuaries, *J. of Environmental Engineering - ASCE*, 138(5), pp.570-577.

부록 3.

Hong-Yeon Cho, Jihee Oh, Kyeongok Kim, and Jae-Seol, Shim, 2013. 4. Outlier detection and missing data filling methods for coastal water temperature data, *Journal of Coastal Research*, Special Issue, No. 65, pp.1898-1903.

부록 4.

Khil-Ha Lee, Hong-Yeon Cho, 2014(예정). Projection of Climate-Induced Future Water Temperature for Aquatic Environment, J. of Environmental Engineering, ASCE)

부 록

부 록 1

어림과 나머지 성분을 이용한 연안 수온자료의 이상자료 감지

한국해양환경공학회지
Journal of the Korean Society
for Marine Environmental Engineering
Vol. 15, No. 2. pp. 156-162, May 2012

기술보고

어림과 나머지 성분을 이용한 연안 수온자료의 이상자료 감지

조홍연† · 오지희
한국해양연구원 해양환경보전연구부

Outlier Detection of the Coastal Water Temperature Monitoring Data Using the Approximate and Detail Components

Hongyeon Cho† and Jihee Oh
Marine Environment & Conservation Research Department, KORDI, Ansan PO Box 29, Seoul 425-600, Korea

요 약

연안 환경모니터링 사업이 확대되면서 방대하게 축적되어 있는 연안 환경모니터링 자료의 통계적 분석을 위해서는 모니터링 자료에서 빈번하게 발생하는 이상 자료의 감지 · 처리가 우선적으로 필요하다. 본 연구에서는 연안 환경모니터링 자료의 어림성분과 나머지(또는 잔차)성분을 이용한 이상자료 진단기법을 제안하였다. 주기함수를 이용한 조화분석 방법과 국지 회귀함수추정 방법을 이용하여 각각 어림성분과 나머지성분을 추출한 후, 추출된 나머지성분 자료에 범용적인 Grubbs 검정기법 및 수정표본점수기법을 적용하여 이상자료를 진단 · 제거한 후 이상자료가 제거된 자료로 재구성하는 방법이다. 제안된 이 기법을 국립수산과학원 실시간어장정보시스템 제공하는 연안 수온 연속 모니터링 자료에 적용한 결과 이상자료가 성공적으로 제거되는 양상을 보이는 것으로 파악되었다.

Abstracts – Outlier detection and treatment process is highly required as the first step for the statistical analysis of the monitoring data having many outliers frequently occurred in the coastal environmental monitoring projects. In this study, the outlier detection method using the approximate and detail (or residual) components of the (raw) data is suggested. The approximate and detail components of the data can be separated by the diverse filtering and smoothing methods. The decomposition of the data is carried out by the harmonic analysis and local regression curve, respectively. Then, the Grubbs' test and modified z-score method widely used to detect outliers in the data are applied to the detail components of the water temperature data. The new data set is reconstructed after removed the outliers detected by these methods. It can be shown that the suggested process is successfully applied to the outlier detection of the coastal water temperature monitoring data provided by the Real-time Information System for Aquaculture Environment, National Fisheries Research and Development Institute (NFRDI).

Keywords: outlier(이상자료), approximations and details(어림과 나머지), water temperature monitoring data(수온 모니터링 자료), Grubbs test(Grubbs 방법), modified z-score method(수정표본점수기법), residual(잔차)

1. 서 론

최근 다양한 관측센서를 이용한 연안환경 모니터링 사업이 활발하게 수행되면서 방대한 환경자료가 축적되어 부족한 자료 또는 제한된 자료를 이용하여 수행 · 분석된 과거의 연구 성과가 재검토 · 재해석되어 새로운 현상 및 특성이 발견되고 있다. 과거의 컴퓨터 연산능력 향상이 기여한 과학기술 발달에 버금가는 수준으로 최근에는 관측기술의 발달로 연속 환경모니터링 자료가 축적되면서 새로운 개념으로 해양과학분야의 기술발전에 기여하고 있다.

그러나 방대한 모니터링 자료의 축적은 DRIP(Data rich, but information poor) 현상을 유발하여 종합적이고 체계적인 자료 분석을 통한 정보추출이 제한받고 있다. 이 제한요소 중의 하나는 이상 자료의 처리문제이다.

이상 자료는 기존의 수동(manual) 방식 또는 휴대용 관측 장비를 이용한 간헐적인 관측에서는 인간의 실수 및 기기 검교정(calibration) 문제 등으로 발생하지만, 관측 장비의 계류에 의한 연속관측 과정에서는 주기적인 센서관리 미흡, 안정적인 전원공급 제한, 센서의 오작동 등의 문제로 빈번하게 발생한다.

관측자료의 개수가 적은 경우에는 확실하게 측정범위를 벗어나는 이상 자료는 간단한 범위 처리과정 및 자료 관리자의 판단을 통하여 제거할 수 있으나, 연속적으로 그리고 매우 짧은 시간간격으로 측정되어 자료의 개수가 기하급수적으로 늘어나는 경우 모든 자료를 하나하나 확인하면서 이상 자료를 수작업으로 제거하는 것은 불가능하게 되어 자동화된 또는 체계화된 처리기법을 필요로 한다.

기본적으로 원(原) 자료(raw data) 분석을 선호하는 전문가는 연속적인 환경 모니터링 자료로부터 정보를 추출하고자 하기 때문에 자료 분석을 위한 전처리과정을 불가피하게 거쳐야 한다. 특히, 대부분의 연속 환경모니터링 자료에서 빈번하게 관찰되는 이상자료를 감지하고 처리하는 과정이 필수적으로 요구된다. 일반적으로 자료를 분석하는 연구자는 각자의 경험과 자료특성을 감안하여 이상자료를 처리하여 왔으나 주관적이고 경험적인 요소가 개입되기 때문에 같은 자료인 경우에도 이상자료를 처리한 자료가 서로 다를 수 있기 때문에 분석결과의 차이가 발생할 수 있다. 따라서 객관적인 측면에서 이상자료를 효과적으로 처리하는 기법에 대한 검토가 요구되고 있다.

본 연구에서는 최근 연안 환경모니터링 사업이 확대되면서 방대하게 축적되어 있는 연안 환경모니터링 자료의 통계적 분석을 위한 전처리 과정중의 하나에 해당하는 이상 자료 감지·처리 기법을 제안하는 것을 목적으로 한다. 환경모니터링 자료 중에서 가장 중요한 인자 중의 하나에 해당하는 수온자료(국립수산과학원 제공자료)를 대상으로 본 연구에서 제안한 기법을 적용하고, 이상 자료 제거 기법을 적용한 경우와 적용하지 않은 경우의 통계정보를 비교 분석하여 전처리 기법 적용 효과를 분석하였다.

2. 이상 자료의 기본

이상 자료(outlier)는 어떤 자료가 그 자료를 제외한 나머지 자료와 일관성이 없어(inconsistent) 보이는 자료 또는 분명하게 다른(distinctly different) 독특한 특성을 가진 자료로 정의되기도 하고(Barnett & Lewis[1994]; Hair et al.[2010]), 비정상적으로(unusually) 크거나 작은 자료, 극한 자료(extreme value)로 정의되기도 한다(Agresti & Franklin[2007]; Martinez & Martinez[2005]). 따라서 이상 자료는 자료의 통계정보를 왜곡할 수 도 있기 때문에 이상 자료에 대한 정량적인 사전 검토가 필요하다. 이상 자료는 잘못된 자료와 특이한 자료로 구분할 수 도 있다. 모두 판단이 필요하지만, 잘못된 자료는 제거하여야 하며, 특이한 자료는 별도로 처리하거나 제외하여 통계분석을 수행할 수 있도록 표기(marking)하여 관리할 필요가 있는 자료이다.

한편 언어를 이용한 정의와 다분어 이상 자료에 대한 통계적 기준도 구체적으로 제시되고 있다. 가장 기본적인 정의는 정규분포 또는 기준이 되는 어떤 분포를 가정하고, 평균(m)과 표준편차(SD)의 함수로 정의되는 영역을 벗어나는 자료로 정의한다. 예를 들면, Hair 등[2010]은 표본의 개수가 80개 정도 또는 그 이하에 해당하는 소표본의 경우와 그 이상에 해당하는 대표본의 경우를 구분하여 다음과 같이 이상자료를 정의하고 있다.

소 표본 : m±2.5(SD) 영역을 벗어나는 자료
대 표본 : m±4.0(SD) 영역을 벗어나는 자료

Grubbs[1950] 및 Dixon[1950] 등도 신뢰수준을 포함한 이상 자료 판단을 위한 각각의 통계기준을 제시하고 있다(Garcia[2012]). 기본적으로 이상 자료의 판단기준은 평균을 중심으로 일정 한계범위를 벗어나는 자료를 이상 자료로 간주하는 개념에 기초하고 있다. Grubbs 방법은 Extreme Studentized Deviate($Max_i(|x_i-\bar{x}|/\sigma_x)$, $\bar{x}$, σ_x=각각 자료 x_i의 평균 및 표준편차) 수치를 한계수치와 비교하여 판단하는 방법이며, Dixon 방법은 자료를 정렬하여 전체 구간에 대한 부분비를 수치를 계산하여 판단하는 방법이다. 수정 표본점수방법은 z-score 계산($z=(x_i-\bar{x})/\sigma_x$)과정에서 표준편차 대신에 MAD(median absolute deviation about the median, $\tilde{x}$) 수치로 계산한 z-score ($z=0.6745\ (x_i-\tilde{x})$/MAD)수치를 이용하여 이상자료를 판단하는 방법으로 개념에 차이가 있다.

3. 이상 자료 감지기법

이상 자료 진단·처리기법은 자료의 종류만큼이나 다양하다. 다양한 분야에서 다양한 이상 자료 진단 방법이 제시되고 있으나, 모든 자료에서 적용되는 방법은 어떤 특정한 자료보다는 정규분포를 따르는 독립적인 자료조건을 충족하는 경우의 진단방법이 유일하며, 가장 활발하게 연구가 추진되어 그 기준도 매우 유사한, 정립된 단계에 해당한다고 할 수 있다. 따라서 특정한 분야의 특정한 자료에 국한되어 사용되는 매우 복잡한 통계도구 및 모형을 이용한 방법보다는 본 연구 분야 또는 특정 연구 분야에서 분석하고자 하는 자료로부터 범용적인 이상 자료 진단기준을 적용할 수 있는 자료를 추출하는 과정을 추가하여, 추출된 자료를 이용하여 이상 자료를 판단하는 기법이 체계적인 접근방법이라고 판단되며, 활용범위의 확장도 가능할 것으로 판단된다.

따라서 본 연구에서 제안하는 방법은 모니터링 자료로부터 어림성분과 나머지 성분을 추출하여, 나머지 성분을 대상으로 범용적으로 이용되는 Grubbs 검정기법(Grubbs[1950]) 등을 이용하여 이상자료를 판단하는 방법이다. 본 연구에서 제안하는 처리 기법은 실질적인 상황과 통계기법을 조합한 방법으로 다음과 같은 단계로 구성되어 있다.

제I단계: 가시적 감지단계(Visual Detection Process)

연안에서 연속적으로 측정되는 환경인자는 각각의 상식적인 또는 제한적인 범위를 가지고 있다. 따라서 개략적인 또는 한정된 범위

†Corresponding author: hycho@kordi.re.kr

제시에 의한 방법으로 터무니없는 이상 자료를 쉽게 진단·제거할 수 있다. 물리적으로 부의미한 값(범위)이 발생(DO 농도 또는 오염물질 농도의 경우 음수가 발생하거나, pH 농도가 0–14 범위를 벗어나는 경우; 수온이 영하 10도 이하 또는 40도 이상인 경우 등)하거나 발생가능성이 거의 없는 경우의 조건을 제시하여 이상 자료를 제거하는 가장 초보적인 단계의 이상 자료 제거 기법이다. 이 방법은 관측 자료를 도시하는 경우, 매우 크거나 작은 값의 이상 자료가 포함되는 경우 도시범위의 확장으로 정상범위가 축소되어 자료변화 양상의 도식적인 판별이 곤란하게 된다.

제2단계: 자료를 어림과 나머지 성분으로 구분하는 과정(Smoothing Process)

시계열 자료에서 바로 이상 자료를 제거하는 연구가 활발하게 수행되고 있으나, 자료가 가지는 구조적인 특성이 관측항목별로 서로 상이하기 때문에 적용에 제한이 따른다. 본 연구에서는 시계열자료가 아닌 이론적으로 IID(independent, identically distributed) 조건을 따르는 자료에 대한 다양한 이상 자료 제거기법이 활발하게 이용되고 있기 때문에 IID 조건에 유사한 자료를 도출하기 위한 과정으로 관측된 시계열 자료를 전체적인 변화양상을 표현하는 어림(approximate, smooth) 성분과 나머지(잔차, residual) 성분으로 구분하기 위한 과정이다. 이 방법의 적용단계에서는 아직 이상 자료가 제거되지 않은 상태이기 때문에 Robust 기법 적용이 필요하다. 본 연구에서는 아래에 제시된 방법을 이용하여 어림 성분과 나머지 성분을 각각 추출하였다.

(1) Robust Smoothing 방법(RLOESS)

어림 성분을 추정하는 방법은 자료 Smoothing 과정으로, 자료의 변화양상을 국지적으로 가중치를 부여하여 적절한 함수곡선으로 맞추어가는 LOESS 또는 LOWESS (locally weighted regression procedure) 방법이 널리 이용되고 있으며, 이상자료의 영향을 줄이기 위한 Robust LOESS, 즉 RLOESS 방법도 있다(Martinez and Martinez[2005]). 본 연구에서는 이상자료의 영향을 줄이기 위한 RLOESS 방법을 이용하여 수온자료의 어림성분을 추정하였다.

따라서 본 연구에서는 이상자료를 적절하게 감지하기 위해서는 이상자료의 영향을 줄이기 위한 모형 매개변수의 Robust 추정이 필요하다(Rousseeuw & Leroy[2003]).

(2) Harmonic Analysis 방법(HA)

조화분석은 조석의 성분분석에 널리 이용되는 방법이나, 체계적으로 주기성분을 고려한 기온 및 수온자료의 어림 성분 추정으로도 널리 제안·이용되고 있다(Cho et al.[2010]). 이 방법은 조석에 의한 영향이 우세하지 않고, 연 변화 및 계절변화 또는 그 이하의 변동성분이 포함되어 있는 환경인자 및 기상인자의 어림 성분 추정에 유용한 방법이다. 본 연구에서는 뚜렷한 연 변화 양상을 가지는 수온 성분의 어림 성분추정으로 이 방법을 이용하였다.

제3단계: 나머지 성분의 이상 자료 감지·제거 과정

제 2단계에서 추출한 나머지 성분을 대상으로, 기존에 제시된 기본적인 이상 자료 감지 기법을 적용하여 이상 자료를 추출하였다. 이상 자료의 제거 여부는 연구자의 경험에 의존하여야 한다. 본 연구에서는 매우 전통적이고 널리 이용되고 있는 Grubbs 진단기법(95% 유의수준)을 이용하여 나머지 성분의 이상 자료를 진단하였다. 이상 자료로 진단된 자료는 모두 제거하였다. 그러나 전체적인 자료의 분포양상을 해석하는 경우에는 큰 문제가 없을 것으로 판단되나, 극치해석 등을 수행하는 경우에는 이상 자료로 진단되어도 특이한 자료인지, 잘못된 자료인지를 판단하여 처리여부를 결정하여야 한다.

4. 이상 자료 감지기법의 적용

4.1 연안 수온자료(실시간 어장정보 시스템 자료)

국립수산과학원에서는 어업활동에 필요한 어장환경정보 제공 및 수산업진흥을 위한 기반자료 구축을 목적으로 연안의 양식어장 및 어장예비가 빈발한 해역에 실시간 해양환경정보(수온, 염분, DO 농도 등) 자동관측시스템을 구축하여 운영하고 있다(국립수산과학원[2012]; 관측지점은 Fig. 1 참조). 본 연구에서는 비교적 장기간의 자료가 가용한 상태에 있는 백령도, 완도(청산), 영덕(거부역) 지점의 수온 자료를 대상으로 본 연구에서 제안한 이상 자료 감지기법을 적용하였다. 관측 자료는 30분 또는 1시간 간격으로 제공되고 있으며, 관측지점에 따라 이상 자료 및 결측자료(missing data), 관측기간이 크게 차이가 나고 있다.

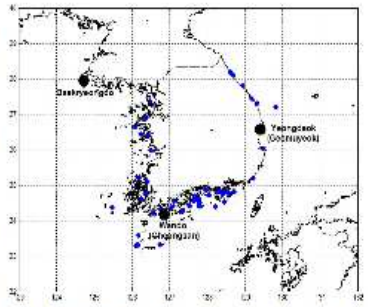


Fig. 1. Location map of the real-time monitoring system of the aqua-culture information (big solid circles : data locations used in this study; Coastline revised from World Vector Shoreline (designed for 1:250,000) data set available through the U.S. National Geophysical Data Center).

4.2 이상 수온자료 감지

국립수산과학원 자료포털에서 다운로드한 자료 그대로를 원자료(raw data)로 가정하고, 제3절에서 제시한 단계를 따라 이상자료를 감지하였다. 제1단계는 가장 기본적인 범위지정에 의한 이상 자료 제거 단계로 범위지정에 의한 이상자료 제거 전·후의 비교 그림이다(Fig. 2 참조). 범위는 하한은 -5 ℃, 상한은 30 ℃ 조건을 사용하였다. 이 범위지정은 간단하게 변경할 수 있으며, 자료의 변동 범위를 감안하여 경험적으로 지정하면 된다. 완도(청산) 지점에서 보이는 바와 같이 정상적인 범위를 크게 벗어나는 잘못된 1–2개 정도의 자료로 인하여 전체적인 자료의 변화범위(-40–160 ℃)가 크게 증가되어 가시적인 수온의 변화양상 파악을 어렵게 하고 있다. 잘못된 자료를 제거하여 자료의 변화구간이 0–30 ℃ 영역으로 제한되는 경우 자료의 변화양상을 시각적으로 쉽게 그리고 보다 뚜렷하게 판단할 수 있다.

범위 지정에 의한 이상 자료를 제거한 후의 과정은 제1단계를 통과한 자료를 이용하여 어림(approximation) 성분과 나머지(잔차) 성분으로 자료를 구분하는 과정이다. 이 과정은 제2단계에서 제시한 RLOESS(Robust LOESS; LOESS=locally weighted regression procedure for fitting a regression curve by smoothing) 방법(국지영역을 지정하는 변수, SPAN=1%)과 조화분석(harmonic analysis, HA) 방법(주기성분은 12개 : 1년 주기부터 1/2년 주기, 1/3년 주기, ..., 1/12년 주기성분까지 이용)을 이용한 자료 변동양상의 근사과정을 통한 어림성분 추출과정과 제1단계를 통과한 자료에서 어림성분을 제외한 나머지 성분추출과정으로 구성된다. 여기서 추출된 나

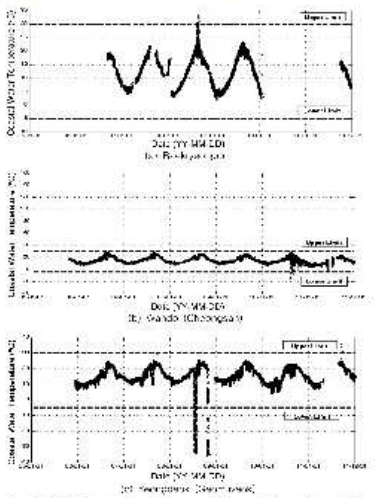


Fig. 2. Outlier detection by the upper and lower limits setting method.

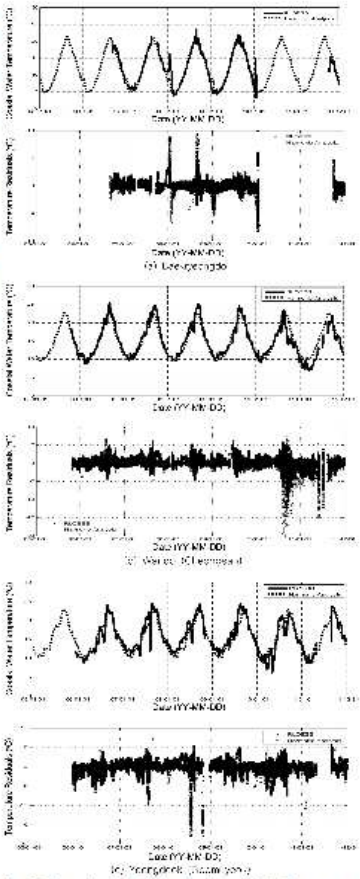


Fig. 3. Time-series plots of the approximate and detail components.

머지 성분을 대상으로 이상 자료 여부를 통계적으로 검정하였다(제3단계). 연안의 수온 변화는 조석의 영향이 중요할 수 있기 때문에 조석성분을 포함하여 조화분석을 수행하였으나, 본 연구영역에

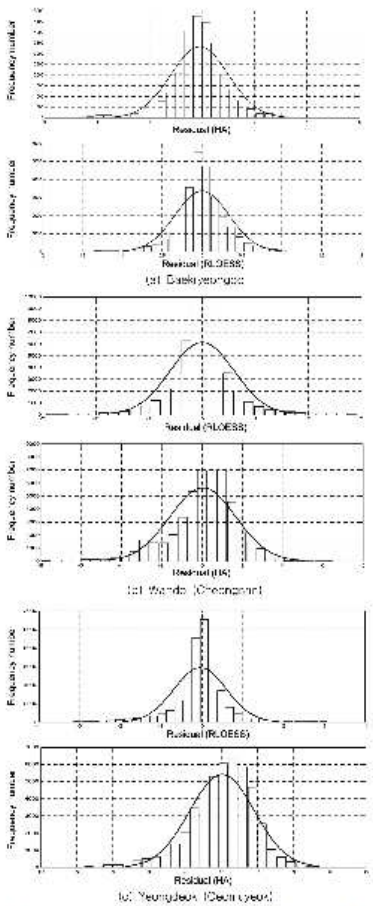


Fig. 4. Histogram of the detail components (Solid line : Normal distribution).

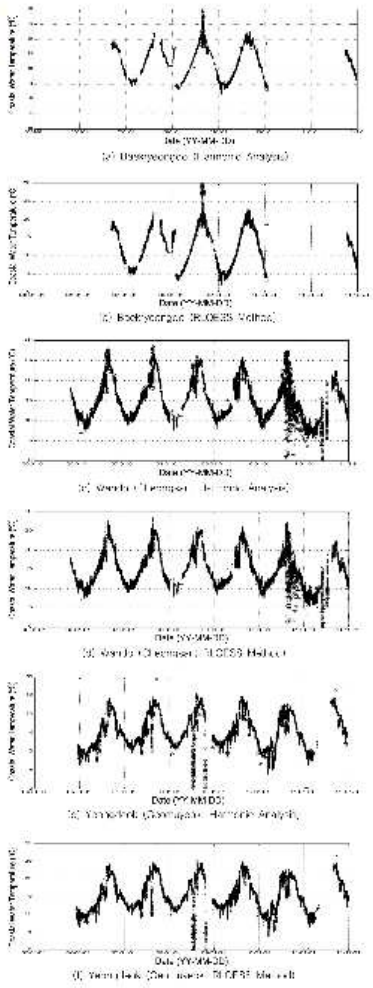


Fig. 5. Outlier and data plots using the harmonic analysis and RLOESS method.

서는 그 영향이 1년 및 1/2년 주기정도의 장주기 성분에 비하여 매우 미미하여 조석성분은 무시하였다.

각각의 방법을 적용하여 추출한 어림 성분과 나머지 성분은 다음과 같다(Fig. 3 참조). 어림성분은 작은 시간규모의 변동성분이 제거되어 보다 평활화(smoothing)된 변화양상을 보이고 있으며, 나머지 성분은 부작위적으로 변동하는 양상을 보이고 있음을 알 수 있다. 나머지 성분은 독립적이고, 정규분포를 따르는 조건을 만족하여야 한다. 제2단계에서 추출된 나머지 성분의 빈도분포와 정규분포를 비교하여 도시하였다(Fig. 4 참조). 그 결과, 나머지 성분이 정규분포와 어느 정도 일치하는 양상을 보이고 있으며, 적절한(robust) 분산매개변수를 선정하는 경우 보다 정규분포에 근사한 분포로 간주할 수 있다고 판단된다. 나머지 성분자료의 분포함수가 반드시 정규분포를 따라야 할 필요는 없고 따르지 않는 경우도 빈번하나, 정규분포에서 크게 벗어나는 경우에는 본 연구에서 사용한 방법에 제한이 따를 것으로 판단된다.

다음은 제3단계에 해당하면서, 가장 핵심이 되는 나머지 성분의 이상자료 감지·제거과정이다. 제2단계의 RLOESS 방법과 HA 방법을 이용하여 추출된 나머지 성분을 Grubbs 기법과 수정 표본점수(z-score) 방법, Dixon 방법 등 범용적으로 이용되는 방법을 이용하여 이상자료 감지에 적용하였으나, 그 감지 결과 차이는 미미하여 Grubbs TEST 적용결과를 대표로 제시하였다. 나머지성분에서 이상자료를 감지·제거하고, 어림 성분을 합하여 이상자료가 제거된 원자료를 재구성하였다. 이상 자료 제거효과는 이상자료를 심볼(◎)로 표시하여 파악할 수 있도록 도시하였다(Fig. 5 참조).

4.3 이상 자료 제거 영향 분석

이상 자료를 제거한 경우, 자료의 통계적인 정보변화를 비교분석하였다. 간단한 통계정보를 정리하였으며, 각각의 단계에 따라 본 연구에서 이용한 수온자료의 평균, 표준편차, 중간값, 평균절대편차 등의 통계정보의 변화를 표에 제시하였다(표 1 참조). 표에서 보이는 바와 같이, 영덕(거부역) 지점의 경우 -5 ℃ 이하의 수온자료가 1,000개 정도 포함되어 있어 표준편차가 제1단계 범위지정 제거과정 이전·이후에 각각 6.96 ℃, 4.50 ℃로 그 차이가 매우 크게 나타났으나, Robust 추정편차에 해당하는 MAD, Median 변화는

Table 1. Basic statistical information changes of the data before and after outlier removals (SD=standard deviation, MAD=median absolute deviation about the median; R=RLOESS Method, H=Harmonic Analysis Method; BOR, AOR=before and after outlier removals, respectively)

(a) Baekryeongdo

Data type	size	mean	SD	median	MAD
raw data	28,217	12.28	5.09	12.03	4.39
data after step 1	28,209	12.27	5.08	12.02	4.39
residual (R/BOR)	28,209	0.01	0.61	0.00	0.29
residual (H/BOR)	28,209	0.00	1.37	-0.05	0.85
residual (R/AOR)	27,765	0.00	0.34	0.00	0.23
residual (H/AOR)	27,910	-0.08	1.10	-0.06	0.78
data after step 3 (R)	27,765	12.24	5.01	11.97	4.33
data after step 3 (H)	27,910	12.22	5.06	11.89	4.37

(b) Wando (Cheongsan)

Data type	size	mean	SD	median	MAD
raw data	49,669	14.87	4.63	14.06	3.76
data after step 1	49,634	14.87	4.53	14.07	3.75
residual (R/BOR)	49,634	-0.01	0.91	0.00	0.48
residual (H/BOR)	49,634	0.00	1.78	0.28	1.25
residual (R/AOR)	48,911	0.00	0.60	0.00	0.41
residual (H/AOR)	49,433	0.04	1.64	0.28	1.21
data after step 3(R)	48,911	14.84	4.47	14.04	3.70
data after step 3 (H)	49,433	14.90	4.52	14.10	3.74

(c) Yeongdeok (Geomyeonk)

Data type	size	mean	SD	median	MAD
raw data	48,147	14.83	6.96	14.62	4.30
data after step 1	47,460	15.45	4.50	14.71	3.79
residual (R/BOR)	47,460	-0.09	1.07	0.00	0.50
residual (H/BOR)	47,460	0.00	2.01	0.25	1.43
residual (R/AOR)	46,905	-0.04	0.64	0.00	0.42
residual (H/AOR)	47,245	0.06	1.79	0.26	1.36
data after step 3 (R)	46,905	15.48	4.43	14.72	3.75
data after step 3 (H)	47,245	15.50	4.45	14.75	3.76

상대적으로 미미하였다. 이상자료의 개수가 많지 않은 경우, 완도(청산) 및 백령도 지점의 경우 및 범위지정에 의한 이상자료 제거(제1단계) 과정 이후에는 전체적인 통계정보 변화에 미치는 영향은 크지 않은 것으로 파악되었다. 이상자료 제거를 통한 표준편차 및 MAD 감소경향은 미미한 수준이나 예상할 수 있는 바와 같이 감소하는 경향을 보였다. 한편 HA 방법에 의한 방법은 RLOESS 방법보다 큰 표준편차를 보이고 있는 것으로 추정되었으나, 이는 각각의 방법의 영향보다는 기법의 적용을 위한 매개변수(근사성분의 개수 및 Robust 추정구간의 범위)의 영향으로 판단된다.

이상자료가 적절하게 제거되었는가하는 정량적인 판단은 곤란하지만, 이러한 판단이 통계적인 의미의 "가장 그럴듯한(most likely)" 진단 관점에서 보면, 자료도시에서 보이는 눈에 거슬리는 이상자료는 상당부분 제거되어 본 연구에서 제안한 방법이 성능을 발휘하고 있는 것으로 판단된다. 또한, 이상자료 제거 전·후의 나머지 성분의 표준편차 변화를 보면, 이상자료 제거전보다 이상자료 제거 후에 표준편차가 감소하게 되는 예측가능한 양상을 보이고 있어, 본 연구에서 제시한 이상자료 제거방법의 효과가 타당함을 보여준다고 할 수 있다.

5. 결론 및 제언

본 연구에서는 연안 환경모니터링 자료의 어림성분과 나머지성분을 이용한 이상자료 진단기법을 제안하였다. HA 방법과 RLOESS 방법을 이용하여 어림성분과 나머지성분을 추출한 후, 추출된 나머지성분 자료에 Grubbs 검정기법 및 수정표본점수 방법을 적용하여 나머지성분을 진단·제거한 후 이상 자료가 제거된 자료를 재구성하였다. 제안된 이 기법을 국립수산과학원에서 제공하는 연안의 수온 연속 모니터링 자료에 적용한 결과, 이상자료가 성공적으로 제거되는 양상을 보이는 것으로 파악되었으며, 본 기법의 적용성능이 우수한 것으로 파악되었다. 영덕(거부역) 지점의 경우 -5 ℃ 이하의 수온자료가 1,000개 정도 포함되어 있어 표준편차가 제1단계 범위지정 제거과정 이전·이후에 각각 6.96 ℃, 4.50 ℃로 그 차이가 매우 크게 나타났으나, Robust 추정편차에 해당하는 MAD, Median 편화는 상대적으로 미미하였다. 이상자료의 개수가 많지 않은 경우 완도(청산) 및 백령도 지점의 경우 및 범위지정에 의한 이상자료 제거(제1단계) 과정 이후에는 전체적인 통계정보 변화에 미치는 영향은 크지 않은 것으로 파악되었다. 그러나 이상자료는 통계정보의 편이(bias)나 왜곡을 유발할 수 있기 때문에 연안 모니터링 자료의 통계적인 분석을 위해서는 반드시 검토하여 처리하여야 한다.

통계적인 의미의 "가장 그럴듯한(most likely)" 진단 관점에서 보면, 자료도시에서 보이는 눈에 거슬리는 이상자료는 상당부분 제거되어 본 연구에서 제안한 방법이 성능을 발휘하고 있는 것으로 판단된다. 한편 이상자료와 더불어 통계적인 정보의 편이(bias)를 유발하는 결측자료 보충(missing data filling-in or imputation) 등의 처리기법에 대한 연구도 수행되어져야 할 것으로 판단된다.

감사의 글

본 연구는 한국해양연구원 기본연구사업(PE98743)의 지원을 받아 수행되었습니다. 연구비 지원에 감사드립니다. 또한 본 연구에서 사용한 어장환경정보시스템 자료를 제공해주신 국립수산과학원에 감사드립니다.

참고문헌

[1] 국립수산과학원, 2012, 실시간 어장정보시스템. http://portal.nfrdi.re.kr/risa/.

[2] Agresti, A. and Franklin, C., 2007, Statistics, The Art and Science of Learning from Data, Pearson Education, Inc. pp.693.

[3] Barnett, V. and Lewis, T., 1994, *Outliers in Statistical Data*, Third Edition, John Wiley & Sons, Ltd., Chichester, UK., pp.584.

[4] Cho, H.Y., Suzuki, K. and Nakamura, Y., 2010, Hysteresis loop model for the estimation of the coastal water temperatures, -by using the buoy monitoring data in Mikawa Bay, Japan-, Report of the Port and Airport Research Institute, 49(2), pp.123-153.

[5] Dixon, W.J., 1950, Analysis of Extreme Values, The Annals of Mathematical Statistics, 21(4), pp.488-506.

[6] Garcia, F.A.A., 2010, Tests to identify outliers in data series, http://www.se.mathworks.com/matlabcentral/fileexchange/28501, MATLAB Central File Exchange. Retrieved January 19th, 2012.

[7] Grubbs, F.E., 1950, Sample Criteria for Testing Outlying Observations, The Annals of Mathematical Statistics, 21(1), pp.27-58.

[8] Hair, J.F. Jr., Black, W.C., Babin, B.J. and Anderson, R.E., 2010, *Multivariate Data Analysis; A Global Perspective*, Seventh Edition, Chapter 2, Pearson Education, Inc., New Jersey, USA., pp.800.

[9] Martinez, W.L. and Martinez, A.R., 2005, Exploratory Data Analysis with MATLAB, Computer Science and Data Analysis Series, Chapman & Hall/CRC. pp.405.

[10] Rousseeuw, P.J. and Leroy, A.M., 2003, Robust Regression and Outlier Detection, John Wiley & Sons. pp.329.

2012년 1월 19일 원고접수
2012년 4월 3일 심사수정일자
2012년 4월 10일 게재확정일자

부 록 2

Development of an Air - Water Temperature Relationship Model to Predict Climate-Induced Future Water Temperature in Estuaries

Development of an Air–Water Temperature Relationship Model to Predict Climate-Induced Future Water Temperature in Estuaries

Hong-Yeon Cho[1] and Khil-Ha Lee[2]

Abstract: To predict climate-induced change in aquatic environments, it is necessary to understand the thermal constraints of various fish species and to understand the timing of current and projected coastal temperatures. This paper presents a newly developed model of the relationship between air and water temperature that was constructed on the basis of harmonic analysis. The model is novel because it requires only a single variable (air temperature) to predict water temperature and captures the hysteresis patterns of the rising and falling limbs and their historic memories. The model was calibrated and validated with data collected from monitoring buoys in Mikawa Bay, Japan between 2005 and 2009. The model validation showed a good performance with a root mean squared error (RMSE) in the range of 0.8–1.0°C. It is especially encouraging that the suggested model can predict water temperature with a reasonable level of accuracy once an acceptable relationship between air temperature and water temperature has been constructed from previously measured data. **DOI: 10.1061/(ASCE)EE.1943-7870.0000499.** © *2012 American Society of Civil Engineers.*

CE Database subject headings: Temperature effects; Water temperature; Climate change; Hysteresis; Estuaries.

Author keywords: Air temperature; Water temperature; Climate change; Harmonic analysis; Hysteresis.

Introduction

Coastal bays and estuaries act as transition zones between upland streams, and coastal oceans are important nurseries and feeding grounds for a large number of marine species (Struyf et al. 2004; Bilgili et al. 2005). It is important to understand how water temperature affects the dissolved oxygen (DO) level in estuarine habitats. Saturated DO levels are lowest at higher water temperatures, which usually occur during the summer (Lee and Lwiza 2008). Critically low DO levels were recently observed in some estuaries and bays and place certain species at risk. Water temperature varies with air temperature, and both have seasonal and diurnal patterns.

Temperature also controls the rates at which chemical reactions and biological processes (such as metabolism and growth) take place. Temperature and salinity variations combine to determine the density of sea water, and this density greatly influences vertical water movement with consequent changes in chemical and biological processes within the water column and the surface sediment layer (Lalli and Parsons 1997). Water temperature partly determines the concentration of dissolved gases in sea water; these gases include oxygen and carbon dioxide, which are profoundly linked with biological processes. Temperature is also one of the most important a-biotic factors influencing the distribution of marine species (Lalli and Parsons 1997). Temperature controls the rate of metabolic and reproductive activities and determines whether fish species can survive (Murphy 2009). It is an essential factor that must be considered in any study involving heat budget computation in semienclosed bays, shallow estuarine zones, and marginal seas (Hsu 1988).

Anthropogenic global warming will influence the thermal dynamics of aquatic environments, and climate changes may impact the organisms living in those aquatic environments [Schirmer and Schuchardt 2001; Struyf et al. 2004; Intergovernmental Panel on Climate Change (IPCC) 2007; Tibby and Tiller 2007]. To project aquatic habitat changes under future climate change conditions, it is necessary to understand the thermal constraints of various aquatic species and to predict the timing of changes in current and projected inland stream and estuarine temperatures. However, a lack of available data often limits our ability to estimate long-term water temperature variation.

In response to a buildup of greenhouse gases, the effect of air temperature on water temperature has been observed. Some efforts have already been made to build a model of air–water temperature relationship in an inland stream (Mohseni et al. 1998, 1999, 2002; Mohseni and Stefan 1999; Morril et al. 2005; Cho et al. 2007; Lee 2007). In fact, a complete heat transfer equation and numerical models for the air–water interface have been developed to estimate water temperature as a function of climate variables (Stefan and Sinokrot 1993). However, these equations and models require large data sets, including information about solar radiation or sunlight time, rainfall and snowfall, wind velocity and direction, humidity, air pressure, topography, altitude and longitude, and the air and water temperatures of the simulation area. Although these mechanism-based models may be more accurate than statistical models, the accuracy of the simulation results depends highly on input data conditions and the condition of the complete input data set for the model run, both of which are rarely satisfactory. Because of the complexity of the complete heat transfer equation, simple regression methods [e.g., a simple linear regression model or the nonlinear S-shape curve (logistic) model] were the dominant method of relating air temperature to water temperature in stream/inland stream water sources (Mohseni et al. 1998; Mohseni and Stefan 1999; Morril et al. 2005; Benyahya et al. 2007).

However, most studies seeking to estimate water temperatures in bay and coastal areas are limited because of insufficient continuous monitoring data. The variations between air and surface water temperatures in inland stream and coastal areas, however, do have very similar patterns on the whole (Knauss 1978; Berner and Berner 1987). This study draws on previous research on inland streams to present a newly developed model of the air–water temperature relationship for coastal and estuarine environments. The model uses a simple regression approach and is based on the traditional method of harmonic analysis. Data collected from monitoring buoys in the Mikawa Bay in Japan between 2005 and 2009 were used to calibrate and validate the new model. The results were then compared with the error bounds for a simple linear regression model.

[1]KORDI, 1270 Sa2-dong, Sangnok-gu, Ansan-si, Kyunggi-do, 426-744, S. Korea. E-mail: hycho@kordi.re.kr

[2]Civil Engineering, Daegu Univ. (DU), Jillyang, Gyeongsan, Gyeongbuk 712-714, S. Korea (corresponding author). E-mail: klee@daegu.ac.kr

Note. This manuscript was submitted on February 15, 2011; approved on September 22, 2011; published online on September 26, 2011. Discussion period open until October 1, 2012; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Environmental Engineering***, Vol. 138, No. 5, May 1, 2012. ©ASCE, ISSN 0733-9372/2012)5-570-577/$25.00.

Materials

Data Used for the Study

The temperature data used for this study were recorded between July 2005 and June 2009. (48 months) and consists of 1,461 carefully screened daily values. The air and water temperature data were provided by the Fishery Experiment Station, Aichi Prefecture, Japan (2010). Secondary air temperature data (included to ensure an accurate analysis) was provided by the Japan Meteorological Agency (JMA 2010). A summary of the site information, including the mean, maximum, and minimum temperature at each station used in this investigation, is presented in Table 1. The measured stations are shown in Fig. 1.

The study sites were selected on the basis of data completeness and reliability. If no daily observations were available, values were estimated by using the stochastic linear regression and Bayesian methods (Akaike 1980; Suzuki et al. 2005). The measured data were then checked for integrity, quality, and reasonableness. The data quality and integrity check were completed for all locations and were then followed by information from precedent studies (Little and Rubin 2002; Barnett and Lewis 1994).

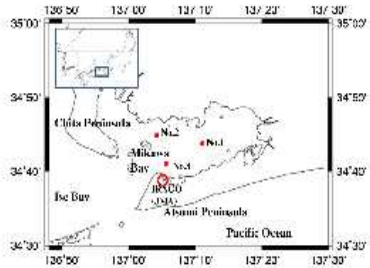


Fig. 1. Location of the Coastal Monitoring Buoy and Irago Weather Monitoring Station (JMA) in Mikawa Bay (Ise Bay) (Buoy 1: 34°44′36″N, 137°13′13″E, depth = 10.1 m; Buoy 2: 34°44′42″N, 137°4′19″E, depth = 10.0 m; Buoy 3: 34°40′30″N, 137°5′49″E, depth = 13.7 m; Irago JMA Station: 34°37′42″N, 137°5′36″E, elevation: +6.2 m)

Preliminary Study

A preliminary study was undertaken to observe the general characteristics and/or seasonal pattern of the air and water temperature. In particular, a relative comparison of the air temperature recorded by the three buoys and the JMA station was performed to ensure data reliability. Fig. 2 presents a scatter plot of the air temperature measurements recorded by Buoy 1 and the JMA station and shows that the coefficient of determination was approximately 0.99. This finding implies that the air temperature measurements from both reliable and interchangeable to some extent. As Fig. 2 shows, there were no noticeable outliers. A temperature data was frequently missing for various reasons and had to be estimated by neighborhood or extrapolation.

Fig. 3 shows the time course of the temperature data for Buoy 1. The temperature data show an annual pattern of variation and thermal stratification between the surface and bottom temperatures are evident. The data also suggest that there is a time lag between changes in the air and water temperatures, but this time lag is not clearly shown in Fig. 3. The time lag is the delayed response of water temperature to air temperature changes attributed to the thermal inertia of water, and it is a function of the average water

Table 1. Basic Statistical Information about the JMA and Buoy Monitoring Data

Items		Mean	Standard deviation	Maximum temperature	Minimum temperature	Number of the missing data
AT (JMA Irago)		16.34	7.68	30.6	0.5	3
Buoy 1	AT	16.69	7.71	30.3	1.6	13
	Surface WT	16.94	7.37	29.9	4.2	26
	Bottom WT	15.93	6.09	26.9	4.3	45
Buoy 2	AT	16.34	7.67	29.9	1.4	13
	Surface WT	17.30	6.93	29.8	5.1	13
	Bottom WT	16.39	5.79	26.7	5.4	13
Buoy 3	AT	16.45	7.54	29.7	1.4	41
	Surface WT	16.85	6.77	29.3	5.3	66
	Bottom WT	16.24	5.71	25.6	5.5	47

Note: These values were calibrated from data collected between July 2005 and June 2009 (total data numbers = 1,461); relatively little data was missing (less than 5%); AT denotes air temperature (°C) and WT denotes water temperature (°C).

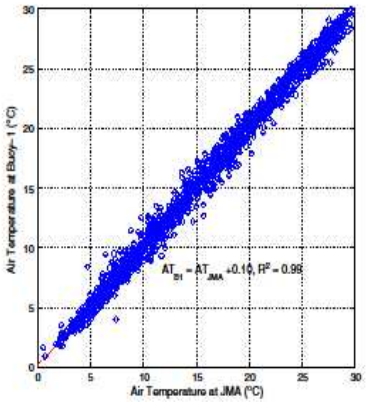


Fig. 2. A scatter plot of the air temperature data recorded by buoys and the JMA stations

depth. Hence, time lag may need to be incorporated into the regression analysis for a certain time scale (e.g., daily) (Stefan and Preud'Homme 1993; Mohseni and Stefan 1999).

To identify the time-memory (or history) of the relationship between air temperature and water temperature, an autocorrelation analysis was performed, and the corresponding results are shown in Fig. 4. However, the autocorrelation coefficient shows a slowly decreasing trend as the time lag increases, which implies that the memory effect is not negligible. Accordingly the memory effects should be included in the model structure in a suitable manner.

In a bay and estuary, several factors may influence the relationship between air and water temperature; these factor include human use, current temperature, air–water interface, and heat exchange. The air–water temperature relationship may have a seasonal hysteresis with a rising and a falling limb. Fig. 5 shows a scatter plot of water temperature against air temperature at Buoy 1 station. There is a noticeable hysteresis pattern in the air–water temperature data. The starting point of the annual cycle is on the bottom-left, and the cycle moves counterclockwise.

Theoretical Background

Harmonic Analysis

First, a regression curve was independently constructed to have the best fit for both the air and water temperature data and to meet the following condition:

$$O = \text{Min}\left\{\sum_{i=1}^{N}[T_E(t_i) - T_O(t_i)]^2\right\} \quad (1)$$

where T_E = estimated temperature; and T_O = observed temperature. The regression curve consists of a summation of the harmonic functions that have different frequency terms

$$T_O(t_i) = T_E(t_i) + \varepsilon_i(t_i)$$
$$T_E(t_i) = A_o + \sum_{m=1}^{M}[A_m\cos(\omega_m t_i) + B_m\sin(\omega_m t_i)] \quad (2)$$

where A_o = mean values of the time-series data (μ_T); A_m and B_m = harmonic coefficients of the order m; M = maximum order of the frequency function in the harmonic analysis; ε_i = residual term (error terms); Y = number of days per year (365 in a common year and 366 in a leap year; 365.25 days were used for more than 1 year); and $T(t_i)$ = temperature data of the i^{th} day, which are composed of $T_1, T_2, \ldots, T_N$ data in which N is the total number of measurements for N days.

The frequency function can then be expressed as follows:

$$\omega_m = 2\pi m/Y \quad (3)$$

where m = the order of the harmonic analysis, and the optimal order (model selection) can be determined from Akaike's Information Criterion (AIC) values (Kitagawa 2005). The multiyear (n-year) periodic function [e.g., $2\pi m/(nY)$] could be included with ease

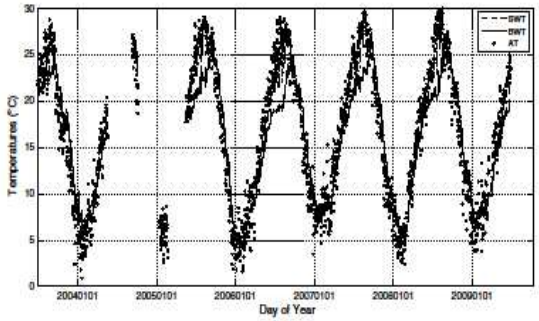


Fig. 3. A time course of the temperature data for Buoy 1 station

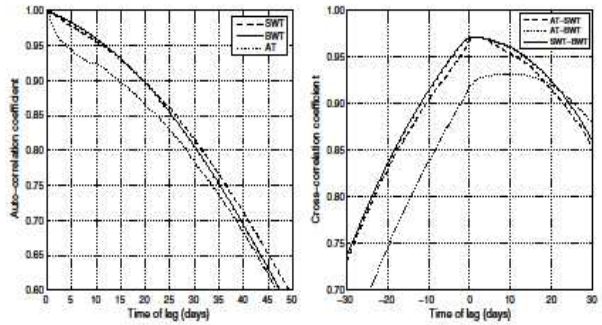


Fig. 4. Autocorrelation coefficients to identify the time-memory (or history)

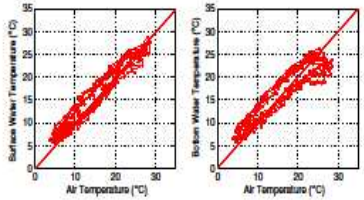


Fig. 5. A scatter plot of weekly mean water temperature against weekly mean air temperature, showing the hysteresis pattern between air–water temperature data

in the harmonic analysis to incorporate long-term frequencies over 1 year.

The harmonic coefficients with order M are estimated in such a way that the difference between the time-series for measured and modeled temperature is minimized. In general, the optimal order of air temperature was higher than that of water temperature for the study site. Table 2 shows the analyses of amplitude and the phase of the harmonic components as a function of order. The amplitude rapidly decreased until the order of 3–5 and slowly decreased after that point as harmonic order increased. The optimal order for the regression curve was determined on the basis of the lowest AIC values that are widely used as model selection criteria (Kitagawa 2005).

Raw data with no harmonic analysis has a higher fluctuation, and the data may make it difficult to analyze overall variation, including peak time, gradient, and time lag. Harmonic analysis is a data smoothing process and therefore, makes an analysis of the overall variation easier.

Air–Water Temperature Relationship

The independent regression curves for air and water temperature data do not provide any relationship between the two variables. Hence, it is desirable to establish a function-type relationship between the air and water temperature data.

As mentioned previously, an air–water relationship can show hysteresis in the rising and falling limbs. Fig. 5 presents scatter plots that shows a hysteresis loop between air and water temperature data for each buoy. Hence, a single valued relationship could not be realistic and sufficient, regardless of its nonlinearity (Lapshin 1995; Cruz-Hernandez and Hayward 2001).

To reflect the hysteresis effect in the model, the following parametric form was introduced as part of this study:

$$[T_w(t), T_A(t)] = [f_A(t), f_w(t)] = [HA(t), HW(t)] \quad (4)$$

where $f_A(t)$ and $f_w(t)$ = best-fitting functions (harmonic function of order M in this study) of the air and water temperature data, respectively. The harmonic functions $HA(t)$ and $HW(t)$ can be expressed as follows:

$$HA(t) - \mu_A^C = \sum_{m=1}^{M}[CA_m^C\cos(\omega_m t) + SA_m^C\sin(\omega_m t)] \quad (5)$$

$$HW(t) - \mu_W^C = \sum_{m=1}^{M}[CW_m^C\cos(\omega_m t) + SW_m^C\sin(\omega_m t)] \quad (6)$$

where μ_A and μ_W = mean air and water temperatures, respectively; CA_m and SA_m = amplitude of the harmonic function of order m on air temperature; and CW_m and SW_m = amplitude of the harmonic function of order m on water temperature. The superscript C denotes the calibration stage.

The parametric form in Eqs. (5) and (6) is advantageous in constructing a variety of different shapes because the order of the harmonic function controls the loop shape. The simplest shape is the ellipse-type loop with Order 1 (see Fig. 6) and the loop shape is usually more complex as the order increases. Order 0 denotes the center point of the loop that corresponds to the location of the mean values.

The primary goal of this study was to estimate water temperature (unknown) by using air temperature (known). However, future water temperature is estimated from future air temperature derived from the scenario- based climate model, and no field record of

Table 2. Amplitude and Phase of the Harmonic Components as a Function of Order

	Order	JMA	Buoy 1			Buoy 2			Buoy 3		
		AT	AT	SWT	BWT	AT	SWT	BWT	AT	SWT	BWT
Amplitude (°C)	0	16.45	16.46	16.88	15.77	16.31	17.30	16.47	16.53	17.01	16.38
	1	10.48	10.45	10.20	8.32	10.39	9.52	7.83	10.18	9.25	7.73
	2	0.78	0.85	0.85	1.76	0.85	0.59	1.52	0.77	0.77	1.71
	3	0.35	0.25	0.11	0.47	0.31	0.17	0.50	0.29	0.11	0.41
	4	0.26	0.23	0.22	0.10	0.27	0.25	0.17	0.24	0.22	0.14
	5	0.23	0.16	0.14	0.18	0.23	0.10	0.16	0.23	0.11	0.17
	6	0.17	0.20	0.12	0.09	0.19	0.16	0.12	0.21	0.11	0.11
	7	0.12	0.12	0.16	0.06	0.11	0.19	0.11	0.08	0.14	0.13
	8	0.12	0.15	0.11	0.06	0.19	0.20	0.08	0.16	0.12	0.04
	9	0.13	0.18	0.15	0.18	0.18	0.14	0.12	0.15	0.14	0.09
	10	0.18	0.17	0.08	0.05	0.16	0.08	0.02	0.16	0.06	0.05
Phase (days)	0	—	—	—	—	—	—	—	—	—	—
	1	33.30	34.84	42.75	50.08	35.31	44.66	51.04	36.48	47.10	52.40
	2	−10.92	−6.36	10.27	20.32	−7.07	10.73	21.65	−3.12	18.73	23.46
	3	−13.89	−13.47	1.11	−28.35	−8.20	−4.84	30.18	−13.13	−0.77	−28.80
	4	−3.31	0.42	11.12	−13.25	1.12	6.59	−21.76	−0.88	7.54	20.85
	5	1.08	−6.39	−4.92	7.22	−3.21	−7.89	3.20	−0.76	1.70	2.02
	6	−1.75	−0.74	−1.92	−14.55	−3.92	−5.31	3.32	−1.95	0.36	9.52
	7	−0.36	4.19	10.11	3.62	6.36	11.62	5.23	5.50	12.57	10.47
	8	−10.93	−9.75	−8.26	−2.23	−9.53	−7.12	−8.74	−9.41	−3.67	−6.43
	9	−2.43	−3.47	2.53	5.41	−2.58	2.04	4.82	−2.48	3.76	0.60
	10	4.89	3.44	8.36	8.93	3.34	4.95	−5.17	3.31	7.65	−0.73

Note: AT, SWT, and BWT indicate the air, surface, and bottom water temperatures, respectively; order "0" indicates the mean value.

water temperature for the future was available. It was, therefore, impossible to derive any harmonic coefficients of water temperature for the Eq. (6) in reality. Therefore, the suggested approach is based on the assumption that the ratio and/or difference of the harmonic coefficients between the air and water temperatures remain constant in the future. Eventually Eqs. (5) and (6) were combined to offer the following relationship between the two temperatures:

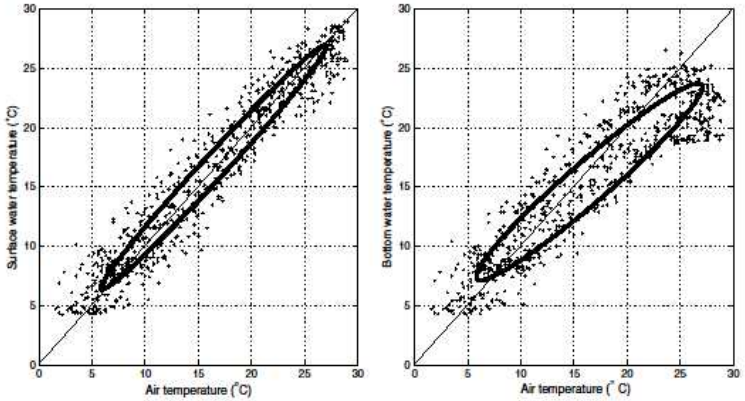


Fig. 6. The simplest shape of the ellipse-type loop with Order 1

$$HA(t) - \mu_A^V = \sum_{m=1}^{M} [CA_m^V \cos(\omega_m t) + SA_m^V \sin(\omega_m t)]$$

$$HW(t) - r_0\mu_A^V = \sum_{m=1}^{M} [cr_m CA_m^V \cos(\omega_m t) + sr_m SA_m^V \sin(\omega_m t)] \quad \text{or}$$

$$HW(t) - d_0\mu_A^V = \sum_{m=1}^{M} [cd_m CA_m^V \cos(\omega_m t) + sd_m SA_m^V \sin(\omega_m t)] \quad (7)$$

where $r_0 = \mu_W^C/\mu_A^C$; $cr_m = CW_m^C/CA_m^C$; $sr_m = SW_m^C/SA_m^C$ [subsequently called r_m (ratio)-type]; $d_0 = \mu_W^C - \mu_A^C$; $cd_m = CW_m^C - CA_m^C$; $sd_m = SW_m^C - SA_m^C$ [subsequently called d_m (difference)-type]; and subscript m = order of the harmonic functions. The superscript V denotes the validation stage. The parameters CA_m and SA_m are separately determined from the observed data set in the validation and calibration stages, and there are different values for the validation and calibration stages unless otherwise stated.

Results

The data were divided into two groups: data from July 2005 to June 2007 (24 months) were used for calibration, data from July 2007 to June 2009 (24 months) were used for validation. In particular, the time lag was investigated. Time lags between air and surface water temperatures changes are much larger in the falling limb, whereas time lags between surface and bottom water temperatures changes are much larger in the rising limb. The time lags between surface and bottom water temperatures changes were more than 30 days. Results also indicated that most of the heat transferred from the surface was used to build thermal stratification during the warm season.

Table 3. Comparison of RMSE (°C) for the linear regression model and the suggested model

		Calibration stage		Validation stage		
Buoy number	Items	LRM	HLM	LRM	HLM-D	HLM-R
Buoy 1	AT—SWT	1.500	0.972	1.415	0.937	0.864
	AT—BWT	2.062	0.972	1.838	0.856	1.418
Buoy 2	AT—SWT	1.482	0.820	1.360	0.852	0.921
	AT—BWT	1.946	0.874	1.765	0.835	1.240
Buoy 3	AT—SWT	1.532	0.847	1.557	0.860	0.762
	AT—BWT	1.903	0.872	1.794	0.788	1.102

Note: AT, SWT, and BWT indicate the air, surface, and bottom water temperatures, respectively; LRM and HLM indicate the linear regression method and the hysteresis loop model, respectively; D and R refer to the difference- and ratio-type methods, respectively.

Calibration of the Suggested Model

To quantify the efficiency of fit, root mean squared error (RMSE) was applied as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} [T_E(t_i) - T_o(t_i)]^2}{N-1}} \quad (8)$$

where N = number of total data points. Fig. 7 presents a scatter plot of water temperature as estimated by the suggested approach in comparison with field observation. A relative comparison of the suggested approach was made against a simple linear regression. Table 3 shows the basic statistics for the two methods. The RMSE for the suggested method was in the range of 0.79–0.97, whereas that for the linear regression method was in the range of 1.42–2.06. The suggested approach showed better performance, and the RMSE for the new method was approximately 5% less at 30°C.

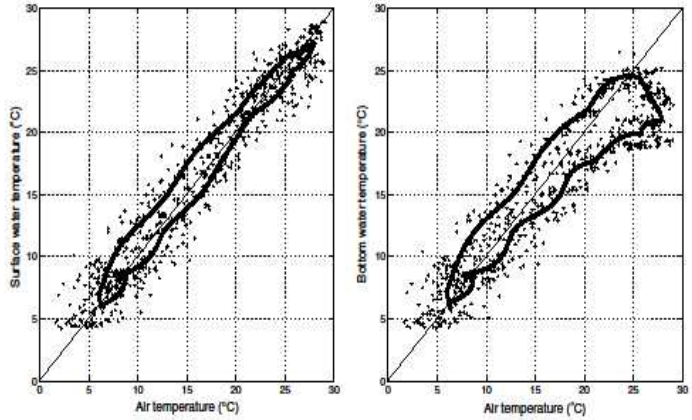


Fig. 7. A scatter plot of water temperature estimated by the suggested approach against field observation: calibration process

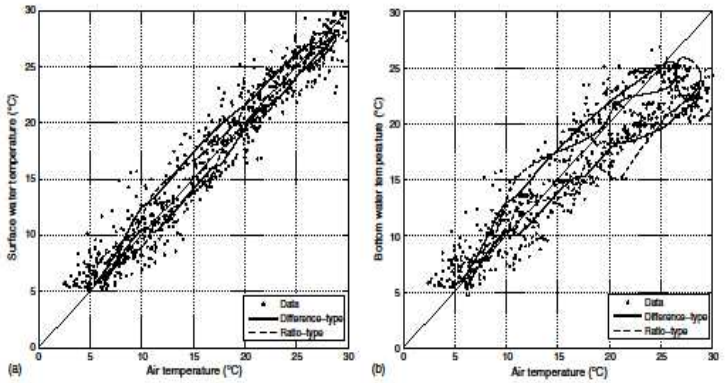


Fig. 8. Scatter plot of (a) surface water temperature and (b) bottom water temperature estimated by the suggested approach against field observation: validation process

Validation of the Suggested Model

The harmonic coefficients derived from the calibration process were used to validate the suggested model. These results are shown in Table 3 and Fig. 8 (air temperature versus surface water temperature; and air temperature versus bottom water temperature). The RMSE of the d_m-type (r_m-type) was in the range of 0.788–0.937 (0.762–1.418), whereas that of the linear model was in the range of 1.360–1.838. Both the r_m-type and d_m-type model showed better performance than did the linear regression model. However, between the two suggested models, the d_m-type model showed better accuracy.

Summary and Conclusions

This study presents a newly developed model of the relationship between air and water temperature. The suggested model, built on the basis of harmonic analysis, is novel because it requires only a single variable (air temperature) as an input to project water temperature. The model captures the hysteresis pattern of the rising and falling limbs and their historic memories. As a result of model validation, RMSE was in the range of 0.8–1.0°C. Therefore, the new model can reproduce water temperature estimations with a reasonable accuracy.

The previous results suggest that the ability to accurately estimate water temperature partly depends on the use of existing data to build an acceptable relationship between air and water temperatures. To some extent, the error may be different for every location of interest, but the method used in this study should work universally. This more realistic description of the air–water temperature relationship is highly desirable and the method suggested in this study could be used to describe the air–water temperature relationship in an inland stream.

Acknowledgments

Hong-Yeon Cho was partly supported by KORDI PE985-01 and KORD PE9853D. We would like to thank all our colleagues at KORDI for their useful and cooperative comments. Also we would like to thank JMA and the Fishery Experiment Station at AICHI Prefecture in Japan for proving the supporting data.

References

Akaike, H. (1980). *Likelihood and Bayes procedure, Bayesian statistics*, J. M. Bernardoet al., ed., University Press, Valencia, Spain.

Barnett, V., and Lewis, T. (1994). *Outliers in statistical data*, 3rd Ed., Wiley, Chichester, UK.

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B. M. J., and Bobee, B. (2007). "A review of statistical water temperature models." *Can. Water Resour. J.*, 32(3), 179–192.

Berner, E. K., and Berner, R. A. (1987). *The global water cycle, geochemistry and environment*, Prentice-Hall, Upper Saddle River, NJ.

Bilgili, A., Proehl, J. A., Lynch, D. R., Smith, K. W., and Swift, M. R. (2005). "Estuary/ocean exchange and tidal mixing in a gulf of Maine estuary: A Lagrangian modeling study." *Estuarine, Coastal Shelf Sci.*, 65(4), 607–624.

Cho, H. Y., Lee, K., Cho, K. J., and Kim, J. S. (2007). "Correlation and hysteresis analysis between air and water temperatures in the coastal zone—Masan Bay." *J. Korean Coastal Ocean Eng.*, 19(3), 213–221 (in Korean).

Cruz-Hernandez, J. M., and Hayward, V. (2001). "Phase control approach to hysteresis reduction." *IEEE Trans. Control Syst. Technol.*, 9(1), 17–26.

Fishery Experiment Station. (2010). (http://www.pref.aichi.jp/suisanshiken) (in Japanese; Feb. 10, 2010).

Hsu, S. A. (1988). *Coastal meteorology*, Academic Press, Waltham, MA.

Intergovernmental Panel on Climate Change (IPCC). (2007). *Climate change 2007: The physical science basis*, Cambridge University Press, Cambridge, UK.

Japan Meteorological Agency (JMA). (2010). (http://www.jma.go.jp) (in Japanese; Feb. 10, 2010).

Kitagawa, G. (2005). "Order determination by AIC." *A premier of the time-series analysis*, Iwanami Press, Tokyo, Japan (in Japanese).

Knauss, J. A. (1978). *Introduction to physical oceanography*, Chapter 3, Prentice-Hall, Upper Saddle River, NJ.

Lalli, C. M., and Parsons, T. M. (1997). *Biological oceanography: An introduction*, Chapter 2, 2nd Ed., Butterworth-Heinemann, Oxford, UK.

Lapshin, R. V. (1995). "Analytical model for the approximation of hysteresis loop and its application to the scanning tunneling microscope." *Rev. Sci. Instrum.*, 66(9), 4718–4730.

Lee, K. H. (2007). "Nonlinear correlation analysis between air and water temperatures in the coastal zone." *J. Korean Coastal Ocean Eng.*, 19(2), 128–135 (in Korean).

Lee, Y. J., and Lwiza, K. M. M. (2008). "Characteristics of bottom dissolved oxygen in Long Island Sound, New York." *Estuarine, Coastal Shelf Sci.*, 76(2), 187–200.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data, Part I*, 2nd Ed., Wiley, Hoboken, NJ.

Mohseni, O., Ericson, T. R., and Stefan, H. (1999). "Sensitivity of stream temperature in the United States to air temperature projected under a global warming scenario." *Water Resour. Res.*, 35(12), 3723–3733.

Mohseni, O., Ericson, T. R., and Stefan, H. (2002). "Upper bounds for stream temperature in the contiguous United States." *J. Environ. Eng.*, 128(1), 4–11.

Mohseni, O., and Stefan, H. (1999). "Stream temperature/air temperature relationship: A physical interpretation." *J. Hydrol. (Amsterdam)*, 218(3-4), 128–141.

Mohseni, O., Stefan, H., and Ericson, T. R. (1998). "A nonlinear regression model for weekly stream temperature." *Water Resour. Res.*, 34(10), 2685–2692.

Morril, J. C., Bales, R. C., and Conklin, M. H. (2005). "Estimating stream temperature from air temperature: Implications for future water quality." *J. Environ. Eng.*, 131(1), 139–146.

Murphy, S. (2009). *General information on temperature*, (http://bcn.boulder.co.us/basin/data/COBWQ/info/Temp.html) (Dec. 20, 2009).

Schirmer, M., and Schuchardt, B. (2001). "Assessing the impact of climate change on the Weser estuary region: An interdisciplinary approach." *Clim. Res.*, 18(1–2), 133–140.

Stefan, H. G., and Preud'Homme, E. B. (1993). "Stream temperature estimation from air temperature." *Water Resour. Bull.*, 29(1), 27–45.

Stefan, H. G., and Sinokrot, B. A. (1993). "Projected global climate change impact on water temperatures in five north central U.S. streams." *Clim. Change*, 24(4), 353–381.

Struyf, E., Damme, S. V., and Meire, P. (2004). "Possible effects of climate change on estuarine nutrient fluxes: A case study in the highly nutrified Schelde estuary (Belgium, The Netherlands)." *Estuarine, Coastal Shelf Sci.*, 60(4), 649–661.

Suzuki, K., Takeda, A., and Hashimoto, N. (2005). "Separation of tidal and sub-tidal currents from intermittent currents data measured at non-fixed routes." *Rep. Port Airport Res. Inst.*, 44(2), 39–56 (in Japanese).

Tibby, J., and Tiller, D. (2007). "Climate-water quality relationship in three Western Victoria (Australia) lakes 1984-2000." *Hydrobiologia*, 591(1), 219–234.

Downloaded 11 May 2012 to 147.46.182.248. Redistribution subject to ASCE license or copyright. Visit http://www.ascelibrary.org

부 록 3

Outlier detection and missing data filling methods for coastal water temperature data

Outlier detection and missing data filling methods for coastal water temperature data

Hong Yeon Cho†, Ji Hee Oh‡, Kyeong Ok Kim† and Jae Seol Shim∞

† Marine Environments and Conservation Research Division, KIOST, Ansan, 425-600, Korea
hycho@kiost.ac
kokim@kiost.ac

‡ Department of Civil & Environmental Engineering, Seoul National University, Seoul, 151-741, Korea
wldl12@snu.ac.kr

∞ Operational Ocean Science and Technology Department, KIOST, Ansan, 425-600, Korea
jsshim@kiost.ac

www.cerf-jcr.org

www.JCRonline.org

ABSTRACT

Cho, H.Y., Oh, J.H., Kim, K.O. and Shim, J.S., 2013. Outlier detection and missing data filling methods for coastal water temperature data. *In:* Conley, D.C., Masselink, G., Russell, P.E. and O'Hare, T.J. (eds.), *Proceedings 12th International Coastal Symposium* (Plymouth, England), *Journal of Coastal Research*, Special Issue No. 65, pp. 1898-1903, ISSN 0749-0208.

Outlier detection and missing data filling (imputation) processes are essential first step in the statistical analysis of coastal monitoring data. Here, we suggest methods for completing these key processes. An outlier detection method that uses approximate and detailed components is suggested. The decomposition of the time-series data is performed by harmonic analysis. Next, the modified z-score method is applied to the residuals (detailed component) to detect outliers. After removing the outliers in the residuals, the filling process for the missing and removed outlier data is conducted by summing the random and the approximate components. Among the environmental monitoring data, this method is applied to the coastal water temperature data. We used hourly interval coastal water temperature data provided by the NFRDI (National Fisheries Research & Development Institute). In these datasets, the dataset of the Yeong-Deok Geomryeok (36.58 °N, 129.40 °E) station, Korea, is only used for this method application. This dataset contains some outliers and missing data. To test the model performance, this method is applied to a daily interval modeling dataset from the HYCOM (Hybrid Coordinate Ocean Model). This method provides reasonable results for outlier detection and for filling in missing data in coastal water temperature datasets.

ADDITIONAL INDEX WORDS: *Outlier, missing data, coastal water temperature, harmonic analysis, modified z-score method, approximation and detail (residual).*

INTRODUCTION

As an important variable for determining climate change, coastal water temperatures (CWT) have been frequently analyzed. To determine the relationships among coastal water temperature and climate change, long-term data must be analyzed. Recently, a variety of coast observational data have been collected due to recent observation technology advancements. However, the accumulation of massive data results in the DRIP (Data Rich, but Information Poor) phenomenon. Thus, information extraction by comprehensive and systematic data analysis has been limited. One limiting factor is the treatment of outliers and missing data. This limitation occurs frequently due to the inadequacy of equipment management, the limited stable power supply, and the malfunctioning of monitoring censors. These limitations affect the reliability of the analyzed results. Consequently, the monitoring data must be preprocessed. Regarding small monitoring datasets, data managers can remove outliers directly with a simple graphical or manual process. However, for massive datasets, an automatic and systematic pre-processing method is necessary.

Generally, the raw data, (i.e., the continuous environmental monitoring data), should undergo a pre-processing step related to data quality control before data analysis(van den Broeck et al., 2005). Principally, outliers detection and removal processes are required as the first step for statistically analyzing monitoring data, as these data include many outliers, which occurred frequently in coastal monitoring projects. Researchers must make judgments on these outliers based on their own experiences and the characteristics of the data; as a result, the outlier detection-removal results may differ. Therefore, objective techniques for effectively detecting and removing outliers are required to produce reproducible results.

In this study, we suggest a method for detecting outliers and for filling in missing data. This method is a statistical analysis preprocess that is applicable to coastal monitoring datasets. Furthermore, this method uses an outlier detection method with approximate and residual data components. The decomposition of the data is conducted with harmonic analysis. Next, the modified z-score method is applied to the residual component to detect data outliers. After removing the outliers, the missing or removed data are filled in. As an example of environmental monitoring data, this method was applied to coastal water temperature. We used hourly coastal water temperature data that were provided by the NFRDI (National Fisheries Research & Development Institute). In these datasets, the dataset of the Yeong-Deok Geomryeok (hereafter YD, 36.58 °N, 129.40 °E) station, Korea, is only used for this method application, which contains some outliers and missing data. Furthermore, this method was applied to a daily modeling dataset from the HYCOM (HYbrid Coordinate Ocean Model).

THE DEFINITION OF OUTLIER

Outliers are defined as inconsistent or distinctly different data (Barnett and Lewis, 1994; Hair Jr. *et al.*, 2010) with unusually small or large and extreme values (Agresti and Franklin, 2007;

DOI: 10.2112/SI65-321.1 received 07 December 2012; accepted 06 March 2013.

Martinez and Martinez, 2005). Thus, data must be quantitatively pre-reviewed because outliers can skew statistical results. In addition, outliers are considered as incorrect or unique data. Judgment is required to determine whether data are incorrect or unique, and incorrect data should be removed. In contrast, unique data should be processed separately or excluded to perform statistical analysis, which is accomplished by flagging the unique data.

Conversely, along with the linguistic definition, specific statistical criteria for outliers have been suggested. The most basic definition, assuming a normal or any standard distribution, excludes data as a function of the mean (m) and standard deviation (SD). For example, Hair *et al*. (2010) separately defined outliers based on the number of samples. A small number of samples corresponded to 80 samples or fewer, and a large number of samples corresponded to more than 80 samples. The definition is as follows:

- Small sample: outside the area of $m \pm 2.5$(SD) and
- Large sample: outside the area of $m \pm 4.0$(SD).

Grubbs (1950) and Dixon *et al.* (1950) suggested that statistical criteria should include the confidence level to detect outliers. These criteria are based on the concept that data that fall outside of a limit range around the average are outliers. The Grubbs outlier detection method uses the ESD (Extreme Studentized Deviate, $\mathrm{Max}\{|(x_i - \bar{x})/\sigma_x|\}$; $\bar{x}, \sigma_x$ = mean and standard deviation of x_i) and the Dixon method uses the computed percentage for the entire interval by sorting the data. The modified z-score method uses the MAD (Median Absolute Deviation about the median, $\tilde{x}$) rather than the standard deviation. Thus, the z-score calculation $(z = (x_i - \bar{x})/\sigma_x)$ becomes $z = 0.6745(x_i - \tilde{x})/\mathrm{MAD}$. The data are regarded as outliers if the absolute values of the data z-score are greater than 3.5 in this study.

METHODS

Outlier Detection

Outlier detection and treatment techniques are as variable as data types. Various outlier detection methods are used in a variety of fields (Choy, 2001; Liu et al., 2004). However, nearly every method is only applicable to independent data that follow a normal distribution. Previous studies have suggested diverse criteria for outlier detection. To use these previous criteria, we divided the time-series water temperature data into two different components, including the approximate components (trends) and the detailed components (residuals). The direct detection of outliers requires a time-series model, such as the ARMA model. However, an outlier detection technique that uses residuals will provide a systematic approach and expand the application range without any time-series model setup.

Therefore, the method suggested in this study decomposes the monitoring data into approximations and residuals. Next, this method detects outliers by applying the modified z-score method to the residual data (Cho and Oh, 2012). Other methods were used to detect outliers, but these results were less sensitive. Certain specified criteria that define the outliers are more sensitive (Cho and Oh, 2012). Thus, the two following steps were used to detect outliers:

- Step 1: Decomposition of approximations and residuals (smoothing process). Research regarding the removal of outliers from the time series data is conducted. However, limitations arose due to the different structural properties between the monitoring data. Various outlier detection methods that can be applied to the data that are theoretically IID (independent and identically distributed) were actively used. Thus, to derive data that follow the IID condition, the data were decomposed into approximate and residual components in this step by harmonic analysis.

  Harmonic analysis is widely applied in tidal component analysis of the tidal elevation and current data, and to estimate the approximation components of the air temperature and water temperature data with systematically periodic components (Emery and Thomson, 2004; Hermance, 2007; Cho & Lee, 2012; Cho *et al.*, 2010). This method is useful for estimating the approximation components of environmental and meteorological variables across annual, seasonal, or lower periodic variations with no tidal effects. In the present study, harmonic analysis was used coastal water temperature data with distinct annual variation. The other smoothing methods, such as moving average or kernel smoothing methods, regression can be used to obtain an approximation component of the data.
- Step 2: Detection and removal of outliers for the residual component. By applying the existing basic outlier detection method, the outliers were extracted from the residual component and decomposed in the first stage. Evaluating outliers relies on the experience of the researchers. In this study, we used the modified z-score method (95 % confidence level), which is more sensitive than the Grubbs method.

Imputation of Missing Data

In addition to outliers, missing data also causes biases regarding statistical information(Little and Rubin, 2002). There are many methods for filling in missing in different fields (Ibanez and Conversi, 2002; Oba et al., 2003; Sehgal et al., 2005). Several of these methods are not applicable to the analysis of data in longitudinal studies, which are correlational research studies that involve repeated observations of the same variables over long periods of time (Twisk and de Vente, 2002). Most statistical packages are only applied to complete datasets without missing data. To overcome this requirement, methods for data imputation have been developed.

As previously mentioned, harmonic analysis is a useful method for estimating data with annual, seasonal, or lower periodic variations. In this study, the missing data were filled in by using harmonic analysis and random number generation using the mean and SD of the residual components. The following steps were used to reconstruct the data:

- Step 1: Decomposition of the data into approximations and residuals (smoothing process). The approximation is the long-term trend of the time series. In contrast, the residual or detailed component is the short-term variation. In this step data were decomposed into approximate and residual components by harmonic analysis.
- Step 2: Reconstruction of the data (imputation process). We obtained the approximation and residual components from the first step. The approximation component was used as the main value, and the residual component was used as a noise component. The noise component is a normal distribution with a mean of zero. In addition, the residual component is normally distributed and is used as a noise component. After calculating the mean and the standard deviations of the residual component, a random number was generated from the mean and standard deviation. Next, the sum of the approximation and the noise (detailed) components were used to fill in the missing data. This method was used to reconstruct the data.

RESULTS

Application to the Monitoring Data

The NFRDI (National Fisheries Research & Development Institute) builds and operates an automatic observation system that provides real-time marine environmental information (including temperature, salinity, dissolved oxygen concentration, etc.) (NFRDI, 2012). Figure 1 shows the monitoring locations. Here, data in YD station were used. These data cover a relatively long time-series (Dec. 2005 - Dec. 2011). The monitoring data were provided every 30 minutes or every hour. The outliers and missing data differed substantially depending on the location and observation period.

The data downloaded from the NFRDI site were available as raw data. In addition, we followed the procedures described in the previous section. In the first step, the raw data were decomposed into approximation and residual components by harmonic analysis (HA). The periodic component of HA was 12. We used a 6-year period. Thus, the components included 6 year and 1/2-year components.

Figure 2 shows the approximation and residual components. The upper panel shows the raw (white circles) and HA data (dashed line). The approximation component was relatively smooth relative to the raw data. The lower panel illustrates the residual component, which oscillates randomly. To apply the next step, the normal distribution assumption must be satisfied. Figure 3 presents a histogram and shows the normal fitting for the residual component. The mean of the residual component was 3.18×10^{-10}, almost zero with a standard deviation of 1.91. The standard deviation changed to 1.33 after outlier removal. This result shows that the relative agrees relatively well with the normal distribution.

The next step was the detection and removal of outliers from the residual component. The residual component follows a normal distribution; thus the modified z-score method was applied to detect the outliers. The detected outliers can be observed in the original data (as shown in Figure 4). The confidence level was 95%.

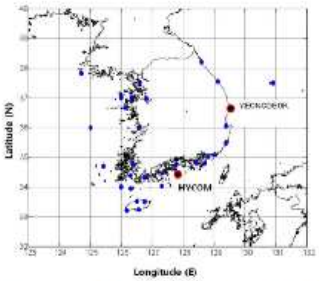


Figure 1. Locations of the real-time monitoring systems for the aqua-culture information (large, solid circle: data location used in this study; coastline revised from the World Vector Shoreline (designed for 1:250,000) data available through the U.S., National Geophysical Data Centre).

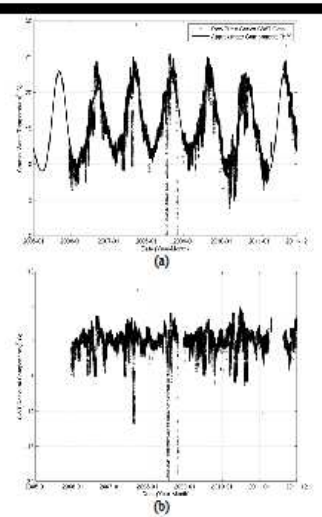


Figure 2. Time series plot of the raw data, approximate components and residual components in YD station. (a) Raw CWT data and approximated components, (b) Residual components.

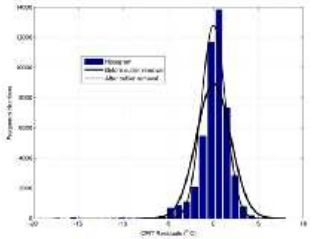


Figure 3. Histogram and normal distributions, $N(0,1.91^2)$ and $N(0,1.33^2)$, fitting of the residual component in YD station.

After removing the outliers, the imputation of missing data was conducted. The first step was to perform HA on the data without outliers. To consider long-term variations, the periodic component included data from the 6-year and 1/2-year components. The

outliers were removed from the data, so that the approximation component followed the original data. To fill in the missing data, the approximation and detailed components are needed. Using the mean and standard deviation of the residual without outliers, random numbers were generated for the detailed component. To fill in the missing data, the sum of the approximation and detailed components was computed. Figure 5 shows the approximation component, the residual component without outliers, and the reconstructed data. Table 1 contains the basic statistical information of the data before and after the outlier removal (BOR/AOR), and after the imputation of the missing data. The mean and median of the step 2 AOR slightly increased relative to the BOR. However, the standard deviation and MAD decreased the AOR. Compared with the original data, the mean and median of the new data (step 3) increased slightly, whereas the standard deviation decreased because of the large negative outliers.

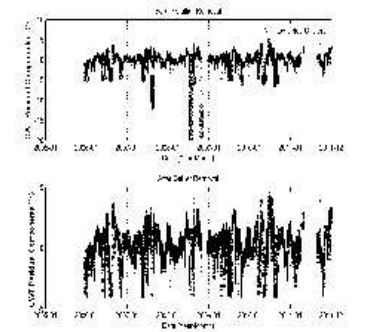


Figure 4. Detected and removed outliers using the modified z-score in YD station.

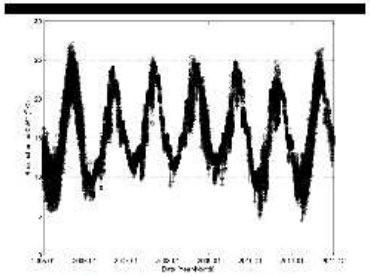


Figure 5. Reconstructed time series data in YD station.

Table 1. Basic statistical information of the data before and after outlier removal and the imputation of the missing data (SD=standard deviation, MAD=median absolute deviation about the median; R=residuals; and BOR and AOR=before and after outlier removal, respectively)

Yeongdeok Geomnyeok (monitoring data)					
Data type	**size**	**mean**	**SD**	**median**	**MAD**
raw data	47,460	15.45	4.50	14.71	3.79
step 2 (R/BOR)	47,460	0.00	1.91	0.25	1.23
step 2 (R/AOR)	45,887	0.22	1.33	0.30	1.02
step 3	52,928	15.64	4.44	14.93	3.79

Application to the Modeling Data

To validate the method, it was applied to the coastal water temperature modeling data. This dataset was provided by the HYCOM (HYbrid Coordinate Ocean model), which can be downloaded from the following site: http://hycom.org/dataserver/glb-analysis (Chassignet *et al.*, 2009). The data collection points are indicated in Figure 1.

The data were collected from the South Sea of Korea at approximately 34.5 °N and 127.8 °E. The data were collected daily between Dec. 2003 and Jul. 2012. To check the performance, some portions of the data (i.e., the high and/or low temperature portions and the intermediate portion) were removed arbitrarily. Next, the following processes were implemented.

Figure 6 (a) shows the raw and arbitrarily removed data. The removed periods included are 1) 01 Nov. 2004 – 31 Jan. 2005, 2) 01 Jul. 2008 – 30 Sep. 2008, 3) 01 Mar. 2009 – 15 Mar. 2009, and 4) 01 Nov. 2010 – 15 Nov. 2010.

In addition, the periodic component of the HA was 18; thus, we used a 9-year period. In Figure 6a, the approximation follows each annual and semi-annual variation shown in Figure 6 (a). Figure 6 (b) shows the residual components of the data. Figure 7 shows the histogram and the normal distribution fitting of the residual components. The summer data in 2008 were removed. Thus, the peak of the approximation is low. In this step, it is important to follow the data trend to detect outliers. Thus, when conducting the HA, the long-term periodic component should be considered. However, this dataset is a modeling dataset. Consequently, few data are detected as outliers.

In the reconstruction step, the periodic component of the HA was 18 from 9 years to 1/2 years. However, the long-term dominant period was 2 years or more due to the missing data at the peak. The optimal long-term period was closely related to the length of the available data record and the relative intensities of the inter-annual and seasonal variations. Figure 8 shows the residual components before and after outlier detections. Figure 9 shows the original and reconstructed data. Figure 10 shows the differences between the original data (not partly removed data) and the reconstructed data. These differences can be regarded of the estimation errors of the missing data filling methods in this study. The mean and standard deviation of the errors are 0.32 and 2.13, respectively.

Table 2 contains the basic statistical information of the data before and after outlier removing outliers and imputing the missing data. Differences were observed when comparing the statistical information between the raw and treated data. However, the statistical information from the raw and reconstructed data was nearly the same. Thus, the reconstructed data well agreed with the raw data.

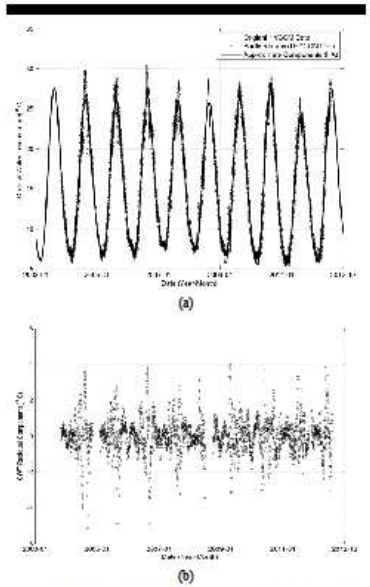


Figure 6. Time series plot of the raw data, approximate and residual components of the HYCOM data. (a) Raw and partly removed data and approximate components, (b) Residual components.

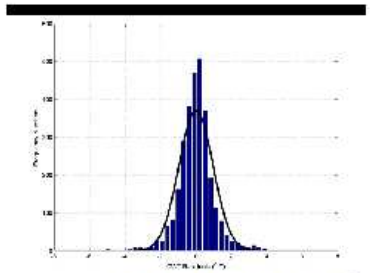


Figure 7. Histogram and normal distribution, $N(-0.02,1.29^2)$, fitting of the residual components of the HYCOM data

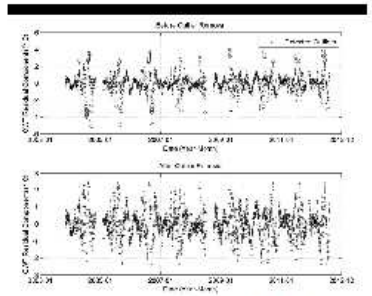


Figure 8. Detected and removed outliers in the HYCOM data

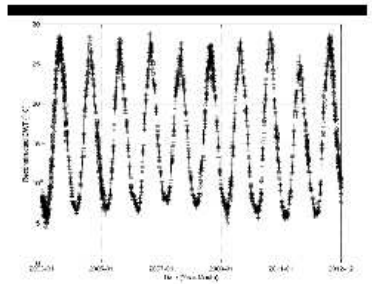


Figure 9. Reconstructed HYCOM time-series data

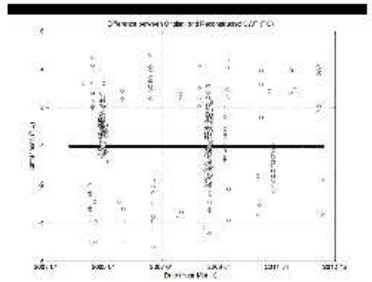


Figure 10. Differences between the original and the reconstructed HYCOM data.

Table 2. Basic statistical information of the data before and after outlier removal and the imputation of the missing data (SD=standard deviation; MAD=median absolute deviation about the median; R=residual; and BOR/AOR=before and after outlier removals, respectively).

Modeling data					
Data type	size	mean	SD	median	MAD
raw data	3,148	15.26	7.00	14.1	6.23
treated data	2,944	15.07	6.93	13.6	6.22
step 2 (R/BOR)	2,944	0.00	1.31	-0.03	1.04
step 2 (R/AOR)	2,938	-0.01	1.30	-0.03	1.04
step 3	3,500	15.25	6.99	14.1	6.28

CONCLUDING REMARKS

In this study, an outlier detection method and a method for filling in missing data (imputation) were used to approximate and detailed (residual) components from coastal environmental monitoring data. By using harmonic analysis, the data were decomposed into approximation and residual components. Next, the outliers were detected for the residual component using the modified z-score, which is more robust to outliers. After removing the outliers, harmonic analysis was applied to the data without outliers. Then, the missing data were filled in with the sum of the approximation and (randomly generated) detailed components, which are a function of the residual components' mean and standard deviation in assumption of the normal distribution.

The suggested method was applied to the coastal water temperature data provided by the NFRDI and the modeling dataset provided by HYCOM. The outliers were successfully removed with this technique. In a comparison of each MAD, median, the changes were relatively minor. However, outliers and missing data caused biases or skewness in the statistical results. Therefore, outliers and missing data should be reviewed and treated before conducting statistical analysis. The HA or any other smoothing methods should be robust and show the intrinsic data trend The longer the missing data interval, the larger the errors of the estimated data. In case of the very long missing data interval, the filling in missing data should be conducted based on the physical-based model not these kinds of statistical models.

From the "most likely" detection of statistical significance perspective, the outliers were successfully removed and the missing data were successfully filled in for this study. Thus, the proposed method is expected to yield high performance.

ACKNOWLEDGEMENT

This study was supported by the following KIOST research projects: PE98823, PE98913 and PE99164.

LITERATURE CITED

Agresti, A. and Franklin, C., 2007. *Statistics, The Art and Science of Learning from Data*. Pearson Education, Inc., 693p.

Barnett, V. and Lewis, T., 1994. *Outliers in Statistical Data, Third Edition*. Chichester, UK: John Wiley & Sons, Ltd., 584p.

Chassignet, E.P., Hurlburt H.E., Metzger E.J., Smedstad O.M., Cummings J., Halliwell G.R., Bleck R., Baraille R., Wallcraft A.J., Lozano C., Tolman H.L., Srinivasan A., Hankin S., Cornillon P., Weisberg R., Barth A., He R., Werner F. and Wilkin J., 2009. U.S. GODAE: Global Ocean Prediction with the HYbrid Coordinate Ocean Model (HYCOM). *Oceanography*, 22(2), 64-75.

Cho, H.Y. and Lee, K.H., 2012. Development of an air-water temperature relationship model to predict climate-induced future water temperature in estuaries. *J. of Environmental Engineering - ASCE*, 138(5), 570-577.

Cho, H.Y., Suzuki, K. and Nakamura, Y., 2010. Hysteresis loop model for the estimation of the coastal water temperatures by using the buoy monitoring data in Mikawa Bay, Japan. *Report of the Port and Airport Research Institute*, 49(2), 123-153.

Cho, H. and Oh J., 2012. Outlier detection of the coastal water temperature monitoring data using the approximate and detailed components. *J. of the Korean Society for Marine Environmental Engineering*, Technical Note, 15(2), 156-162.

Choy, K., 2001. Outlier detection for stationary time series. *Journal of Statistical Planning and Inference*, 99, 111-127.

Dixon, W.J., 1950. Analysis of Extreme Values. *The Annals of Mathematical Statistics*, 21(4), 488-506.

Emery, W.J. and Thomson, R.E., 2004. *Data Analysis Methods in Physical Oceanography*, Second and Revised Edition, Chap. 5, Elsevier.

Garcia, F.A.A., 2010. Tests to identify outliers in data series, http://www.se.mathworks.com/matlabcentral/fileexchange/28501, MATLAB Central File Exchange. Retrieved January 19th, 2012.

Grubbs, F.E., 1950. Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*, 21(1), 27-58.

Hair, J.F. Jr., Black, W.C., Babin, B.J. and Anderson, R.E., 2010. *Multivariate Data Analysis*. A Global Perspective, Seventh Edition, Chapter 2, New Jersey, USA: Pearson Education, Inc., 800p.

Hermance, J.F., 2007. Stabilizing high-order, non-classical harmonic analysis of NDVI data for average annual models by damping model roughness. *International J. of Remote Sensing*, 28(12), 2801-2819.

Ibanez, F. and Conversi, A., 2002. Prediction of missing values and detection of 'exceptional events' in a chronological planktonic series: a single algorithm. *Ecological Modelling*, 154, 9-23.

Little, R.J.A. and Rubin, D.B., 2002. *Statistical Analysis with Missing Data, Second Edition*. Chichester, UK: John Wiley & Sons, Ltd.

Liu, H., Shah, S. and Jiang, W., 2004. On-line outlier detection and data cleaning. *Computers and Chemical Engineering*, 28, 1635-1647.

Martinez, W.L. and Martinez, A.R., 2005. *Exploratory Data Analysis with MATLAB*. Computer Science and Data Analysis Series, Chapman & Hall/CRC, 405p.

NFRDI, 2012. http:// portal.nfrdi.re.kr/risa/

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S., 2003. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16), 2088-2096.

Sehgal, M.S.B., Gondal, I. and Dooley, L.S., 2005. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics*, 21(10), 2417-2423.

Twisk, J. and de Vente, W., 2002. Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology*, 55, 329-337.

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R. and Herbst, K., 2005. Data cleaning: Detecting, diagnosing, and editing data abnormalities. *Plos Medicine*, 2(10):e267, 0966-0970.

부 록 4

Journal of Environmental Engineering

The Projection of Climate-Induced Future Water Temperature for the Aquatic Environment

Journal of Environmental Engineering

The Projection of Climate-Induced Future Water Temperature for the Aquatic Environment

–Manuscript Draft–

Manuscript Number:	EEENG-2075
Full Title:	The Projection of Climate-Induced Future Water Temperature for the Aquatic Environment
Manuscript Region of Origin:	KOREA, REPUBLIC OF
Article Type:	Technical Paper
Manuscript Classifications:	53: Water; 55: Ecology
Keywords:	Water temperature; Air temperature; Climate change; GCM; Global warming scenario
Abstract:	To project the effects of climate-induced change on aquatic environments, it is necessary to know the thermal constraints affecting different fish species and to acquire time series of the current and projected water temperature (WT). A regression between the WT at individual stations and the ambient air temperature (AT) at nearby weather stations could represent the easiest practical method of estimating the WT for an entire region. Assuming that the grid-averaged observations of AT correspond to the AT output from a General Circulation Model (GCM) simulation, this study constructed a regression curve between the observations of the local WT and the concurrent GCM-simulated surface AT, minimizing the difference between the time series of the measured and modeled WT, which implicitly includes downscaling to local conditions. The regression model shows excellent performance in capturing the WT trend in response to the AT of the GCM. The projected WT under the global warming scenario shows a 1.5~2.5 °C increase for the period 2080-2100. The maximum/minimum WT shows an amount of change similar to that of the mean values. The results also predict that the WT will increase during most seasons by 2100 but that only minor changes will occur during the summer.
Corresponding Author:	Khil-Ha Lee, PHD Daegu University(DU) Gyeongsan, Gyeongbuk KOREA, REPUBLIC OF
Corresponding Author E-Mail:	klee@daegu.ac.kr
Order of Authors:	Khil-Ha Lee, PHD
	Hong-Yeon Cho
Suggested Reviewers:	Noname01 None N/A none@none.edu N/A
	Noname02 None N/A none02@none.edu N/A
	noname03 nono N/A none03@none.edu N/A
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Is the article being considered for more than one journal?The Journal of Environmental Engineering does not review manuscripts that are being submitted simultaneously to another organization or ASCE journal for publication.	No
Is this article already published? Material that has been previously published cannot be considered for publication by ASCE. A manuscript that has been published in a conference proceedings may be reviewed for publication only if it has been significantly revised. If you answer YES, please explain.	No
Have all the authors contributed to the study and approved the final version?All authors must have contributed to the study, seen the final draft of the manuscript, and accept responsibility for its contents. It is unethical to list someone as a coauthor who does not want to be associated with the study and who has never seen the manuscript.	Yes
Was an earlier version of the paper previously considered and declined by ASCE?Declined manuscripts are sent through the review process again. If your manuscript has been submitted to us before under a different title, please provide that title in the space provided below. It is our policy to inform an editor that a manuscript has been previously reviewed, even when it has been reviewed by a different Division, Institute, or Council within ASCE.	No
Do your table titles/figure	No

Powered by Editorial Manager® and Preprint Manager® from Aries Systems Corporation

captions cite other sources?If you used a figure/table from another source, written permission for print and online use must be attached in PDF format. Permission letters must state that permission is granted in both forms of media. If you used data from another source to create your own figure/table, the data is adapted and therefore obtaining permission is not required.	
Does your paper exceed 10,000 words? If YES, please provide justification in your cover letter. If you need help estimating word length, see our sizing worksheet at this link: Sizing Worksheet	No
Estimates for color figures in the printed journal begin at $924. Cost increases depend on the number and size of figures. Do you intend for any figure to be printed in color? If YES, how many and which ones? Please provide a total count and also list them by figure number.	No
Is this manuscript a companion to one already submitted/or being submitted? If yes, please note whether this is part I, II, or III. Please make sure all related papers are uploaded on the same day and provide the date of submission, title, and authors of each.	No
Is this manuscript part of a Special Issue? If yes, please provide the Special Issue title and name of the guest editor.	No
Journal of Environmental Engineering Preview is ASCE's initiative to publish author manuscripts online within 72 hours of acceptance and before	Opt-In *(post my uncopyedited manuscript online)*

the final, copyedited version of record is published in print and online. Note: Once the manuscript is posted online, it is considered published. Edits will ONLY be allowed when the corresponding author receives a proof of the composed and copyedited version of the manuscript. Your uncopyedited manuscript will be posted online unless you click the Opt-Out button below. For more information & policy: http://pubs.asce.org/journals/pap/	
To read ASCE's Data Sharing Policy, please click on the "Instructions" link associated with this question.According to this policy, you are required to report on any materials sharing restrictions in your cover letter. Are you restricted from sharing your data & materials? If yes, did you report on these in your cover letter?	No

Powered by Editorial Manager® and Preprint Manager® from Aries Systems Corporation

Cover Letter
Click here to download Cover Letter: Cover Letter.doc

Cover Letter

Jan 01, 2013
Journal of Environmental Engineering-ASCE

To whom it may concern

Please find an enclosed file package of "**The Projection of Climate-Induced Future Water Temperature for the Aquatic Environment** " written by K Lee and H Cho. I would like to have this manuscript reviewed by the JEE-ASCE. We have followed all the instructions for author. It is also noted that it describes original research not submitted or published for somewhere else. Should you need to contact me, please use the information below (email preferred).

Sincerely

Khil-Ha Lee
Daegu University (DU)
Jillyang, Gyengsan, Gyeongbuk 712-714
S. Korea
Tel: +82 53 850 6522
Fax: +82 53 850 6529
Email: klee@deagu.ac.kr

*Manuscript
Click here to download Manuscript: draft 1.0.doc

The Projection of Climate-Induced Future Water Temperature for the Aquatic Environment

Khil-Ha Lee: Corresponding Author
Civil Engineering, Daegu University (DU)
Jillyang, Gyeongsan, Gyeongbuk 712-714
S. Korea
Tel: +82 53 850 6522
Fax: +82 53 850 6520
Email: klee@daegu.ac.kr

Hong-Yeon Cho
KIOST (Korea Institute of Ocean Science and Technology)
Haeanro 787
Ansan-si, Gyeonggido, 426-744
S. Korea
Tel: +82 31 400 6318
Fax: +82 31 400 7868
Email: hycho@kiost.ac

Abstract

To project the effects of climate-induced change on aquatic environments, it is necessary to know the thermal constraints affecting different fish species and to acquire time series of the current and projected water temperature (WT). A regression between the WT at individual stations and the ambient air temperature (AT) at nearby weather stations could represent the easiest practical method of estimating the WT for an entire region. Assuming that the grid-averaged observations of AT correspond to the AT output from a General Circulation Model (GCM) simulation, this study constructed a regression curve between the observations of the local WT and the concurrent GCM-simulated surface AT, minimizing the difference between the time series of the measured and modeled WT, which implicitly includes downscaling to local conditions. The regression model shows excellent performance in capturing the WT trend in response to the AT of the GCM. The projected WT under the global warming scenario shows a 1.5~2.5 °C increase for the period 2080-2100. The maximum/minimum WT shows an amount of change similar to that of the mean values. The results also predict that the WT will increase during most seasons by 2100 but that only minor changes will occur during the summer.

CE Database subject headings: Water temperature, Air temperature, Climate change, GCM, Global warming scenario

Introduction

Water temperature (hereafter abbreviated WT) is an important factor in the aquatic environment. Changes in the WT result in changes in the ecosystem and in water quality, especially in the dissolved oxygen (DO) level, which affects the aquatic biota (Johnson, 1971; Mohseni et al., 1998; Mohseni and Stefan, 1999; Mohseni et al., 1999; Mohseni et al, 2002; Morril et al., 2005; Pilgrim and Stefan, 1995; Stefan and Sinokrot, 1993; Struyf et al., 2004; Webb, 1987). It is therefore important to understand how the WT affects the DO level in aquatic habitats. Over the past century, the average global air temperature (hereafter abbreviated AT) has increased by 1 °C. Anthropogenic global warming has continued and is expected to increase by 1-3 °C by the middle to the end of the century (IPCC, 2007). The effect of the AT on the WT in response to the increase in greenhouse gases has been noted, and efforts have been made to build an AT-WT relationship based on historical records.

In fact, a complete heat transfer equation at the air-water interface may be included for the estimation of the WT as a function of climate variables (Stefan and Sinokrot, 1993). Due to the complexity of the complete heat transfer equation, however, simple regression methods have primarily been used to relate the AT to the WT (Mohseni et al., 1998; Mohseni and Stefan, 1999; Morril et al., 2005).

The outputs of General Circulation Models (GCMs) serve to provide grid-scale projections of the AT. The degree of confidence in the simulations of each model is based primarily on the assumptions and parameterizations used in developing the model (Karl et al., 1990). The impact of climate change is primarily involved with site-specific factors such as water management, the aquatic environment, the ecosystem, public health, agricultural production, disaster management, and coastal management. For this reason, climate change needs to be understood on a local or site-specific basis, and it is very difficult to estimate the impact of climate change on a local scale without an adequate means of relating GCM climate variables to the observed local climate variables.

Regressions between the WT at individual gauging stations and the AT at nearby weather stations can provide an easy practical method to relate the AT to the WT.

Several studies have shown the feasibility of linking large-scale spatial averages to local (station) averages, and the AT from a GCM can reproduce the observed regional and local AT (DeGaetano, 2001; Karl et al., 1990; Oshima et al., 2002; Schubert, 1998; School and Pryor, 2001; Winkler et al., 1997). In this study, we construct projections of the local WT from concurrent GCM simulations of the surface AT. The projected WTs from three study sites are presented and discussed. The study could provide a useful relationship between the local climate and the perturbed simulation in GCMs in future applications.

Materials

Study region and data used for the study

The Korean Peninsula has a moderate climate characterized by distinct wet and dry seasons. The dry season coincides with the northwest wind, which is predominant from October to March. The wet season results from the southeast wind, which carries moisture-laden air from the Pacific Ocean and lasts from May to October, providing 70 % of the annual precipitation. July-August is usually the wettest season. All three sites exhibit typical daily and seasonal variations in moderate temperature trends, with seasonal maxima from May to October and minima from November to April. Estuarine sites generally show less variability in temperature, with a diurnal temperature range (DTR) of approximately 7 °C.

Figure 1 shows the three study locations: Ansan, Masan, and Nakdong (summary in Table 1). The AT-WT data for the three sites are simultaneously collected by the Korea Ocean Research and Development Institute (KORDI). These data were aggregated into weekly averages from 5-min observations for the study. The AT-WT data were collected using an OS-316 CTD sensor (Idronaut S.R.L., Milano, Italy) (approximately 1 m below the water surface for the WT and approximately 2 m above the water surface for the AT). All three stations are within the vicinity of the coastal zones, and they can be affected by the ocean to a certain extent. The Ansan station is in a closed area surrounded by a sea wall, but contact with the ocean occurs on occasion, when the gate to the sea is open. The Masan station is semi-open to the ocean and is partly exposed. The Nakdong station is completely closed off by a sea wall, and there is no contact with the ocean because the sea wall prevents ocean intrusion.

In the aquatic environment of the upland stream, several factors may influence the AT-WT relationship: the distance from the water source, shading, human industrial use, the inflow temperature, and the air-water interface heat exchange (Mohseni et al., 1998; Mohseni and Stefan, 1999; Mohseni et al., 1999; Mohseni et al, 2002; Morril et al., 2005; Pilgrim and Stefan, 1995; Stefan and Sinokrot, 1993). In an estuarine environment, additional oceanic tidal effects may occur with seasonal patterns. There is no reservoir release upstream from the gauging station, but there is some interaction with groundwater inflow into the station. The mechanisms affecting the AT-WT relationship at the estuary are assumed to be similar to those at the upland areas, but it will be shown that this assumption is generally satisfactory. The area of the Korean Peninsula is approximately 100,000 km^2, and a 2-3 ° GCM grid covers the entire area. AT data from 25 stations (**Table 2**) collected by the Korea Meteorological Administration (KMA) are used for the years 1991-2010, and their averaged values are assumed to be a surrogate for GCM-grid values for this study. The measured data were checked for integrity, quality, and reasonableness. The data quality and integrity check were completed for all locations and were then followed by supplemented information from previous studies (Little & Rubin, 2002; Barnett & Lewis, 1994).

Climate change scenarios and GCM data

The GCM data used for the study are freely available (at http://www.cccma.bc.ec.gc.ca), and the climate model simulation was obtained from the Canadian Centre for Climate Modeling and Analysis (CCCma).

The second-generation coupled global climate model (CGCM2) of the CCCma was used to produce ensemble climate change projections using the IPCC SRES A2 and B2 scenarios (IPCC, 2007). The CGCM2 represents the net radiative effect of all greenhouse gases by means of an equivalent CO2 concentration. Both the A2 and B2 scenarios are based on the observed greenhouse gas (GHG) forcing change from 1961 to 1990 provided by the Hadley Centre (hence, the data for 1961-1990 are identical for both scenarios), and daily data for the years 1961-2100 are available. A2 consists of data from a 111-year (1990-2100) ensemble simulation using the provisional IPCC SRES A2 GHG and aerosol forcing scenario, whereas B2 consists of data from a 111-year (1990-2100) ensemble simulation using the provisional IPCC SRES B2 GHG and aerosol forcing scenario. Briefly, the A2 scenario envisions population growth to be 15 billion by 2100 with rather slow economic and technological development, whereas the B2 scenario envisions the population growth to be 10.4 billion by the year 2100, with a more rapidly evolving economy and a greater emphasis on environmental protection. Therefore, B2 is expected to produce lower emissions and less future warming.

Theoretical background

Much research has shown that a linear regression function is insufficient to determine the WT year round because the AT-WT relationship does not usually remain linear at the highest and lowest AT. A harmonic

function has been suggested as an alternative (Cho and Lee, 2012). An air-water relationship can show hysteresis in its rising and falling limbs. Hence, a single-valued relationship cannot be realistic or sufficient, regardless of its nonlinearity (Cho and Lee, 2012). To reflect the hysteresis effect in the model, the following parametric form was introduced as part of this study:

$$[T_w(t), T_A(t)] \sim [f_A(t), f_w(t)] \sim [HA(t), HW(t)], \quad (4)$$

where $f_A(t)$ and $f_w(t)$ are the best-fitting functions (the harmonic function of order M in this study) of the air and water temperature data, respectively. The harmonic functions $HA(t)$ and $HW(t)$ can be expressed as follows:

$$HA(t) - \mu_A^C = \sum_{m=1}^{M} \left\{ CA_m^C \cos(\omega_m t) + SA_m^C \sin(\omega_m t) \right\} \quad (5)$$

$$HW(t) - \mu_W^C = \sum_{m=1}^{M} \left\{ CW_m^C \cos(\omega_m t) + SW_m^C \sin(\omega_m t) \right\}, \quad (6)$$

where μ_A and μ_W denote the mean air and water temperatures, respectively; CA_m and SA_m denote the amplitude of the harmonic function of order m on air temperature; and CW_m and SW_m denote the amplitude of the harmonic function of order m on water temperature. The superscript C denotes the calibration stage. The reader can refer to Cho and Lee (2012) for more details.

Outcomes

Linear relationship between GCM AT and local AT

Because the GCMs provide outputs at a low spatial resolution (up to a few degrees), downscaling to local conditions is essential. Before further analyses were performed, the linear relationship between the GCM grid-scale AT and the local AT was checked. The averaged AT from all 25 stations was considered a surrogate for a GCM-simulated AT. The linearity between all station-averaged values and individual station values was then investigated.

First, an L-moment frequency analysis was performed to check the homogeneity within the GCM grid-scale, and the site was declared discordant if the discordancy index value, D, was beyond the threshold (Hosking, 1997). The spatial grouping of observation sites provides a means of smoothing sampling variations and is the basis for forming homogeneous climate groups on the basis of terrain, vegetation, and climatological characteristics (DeGaetano, 2001). Data screening using the discordancy measure test (Hosking, 1997) was employed to identify those sites from the group of given regions that were discordant with the group as a whole. The sites with large errors in the data set were excluded. It was found that the D values of the selected 25 gauging sites for the AT data were below the critical value, and the data of all 25 gauging stations were treated as homogeneous, consistent with the assumption that the Korean peninsula can be treated as a climatologically homogeneous division.

Figure 2 presents a scatterplot of the 25 station-averaged ATs vs. the representative stations. The coefficient of determination r^2 is approximately 0.98, and all stations provide very good linearity. This

result may imply that the local or site-specific AT can be replaced by the GCM grid-scaled AT with reasonable accuracy in Korea.

Projected WT under global warming scenarios

A detailed analysis of the climate modeling and how these projections were determined is beyond the scope of this paper, and these projections are regarded as the best available at the moment because they are based on research by many experts. However, there may be a gap between the historical data and the GCM outputs, and the various existing statistical transformations are used to ensure that the historical data and GCM outputs have similar statistical properties. The method described by Alcamo et al. (1997) was used for this study. For temperature, the absolute changes between the historical and future GCM time series were added to the measured values as follows:

$$T_{GCM} = T_{his.avg} + (\overline{T}_{GCM.fut} - \overline{T}_{GCM.hist}) \quad (2)$$

in which T_{GCM} is the transformed future temperature, $T_{his.avg}$ is the measured temperature, $\overline{T}_{GCM.fut}$ is the average future GCM temperature and $\overline{T}_{GCM.hist}$ is the average historical GCM temperature. Thus, the average of the transformed GCM temperatures for historical times is the same as that for the measured historical temperatures.

The calibrated model parameters were assumed unchanged for the WT projection, and the physical setting, which includes geometry, wind sheltering, and shade, remains unchanged. Figure 3 shows the WT fitted to the harmonic regression function (hysteresis loop) against the observed WT at the representative station. A statistical evaluation of the relationship between the estimated WT and the observed WT is presented in Figure 4. Figure 4 shows the exceedance probability (EP) of the upper/lower quantiles of the estimated WT for a range of error norms. Three stations show a similar level of statistical accuracies. All three stations are associated with high EP values, ranging from nearly 100% for a low error norm (0.05) to 10-15% for a high error norm (0.2).

Figure 5 presents a time series of the projected annual WT for 2012, 2050 and 2100, and figures 6-7 present 20-year average values at the representative station under the global warming scenario (A2 and B2) simulated by CCCma. In general, the WT in the A2 scenario shows a ~3.0 °C increase for 2080-2100, while the AT in the B2 scenario shows a ~1.5 °C increase.

The main focus of this study is on the impact of climate change on the aquatic environment. In certain cases, the maximum and minimum WTs (see figures 6-7) are more important than the mean annual WT. The boxes in figures 6-7 indicate the averages, and the lower/upper triangles indicate the minimum/maximum values for each time period. For the years 2080-2100, the increases in the maximum/minimum WT are ~1.0 °C for the B2 scenario. The increases in the maximum WT are ~1.5 °C,

but the increases in the minimum WT are ~5 °C for the A2 scenario. The projected WT shows large changes in the maximum/minimum WT compared with the mean WT under the climate change scenarios. Figure 8 presents the annual variation of the AT and WT for the year 2100. The projected WT shows an increasing trend in 2100, but no seasonal patterns are found.

Summary and conclusions

Assuming that the averaged AT from 25 stations in Korea corresponds to the output from a GCM simulation, each site studied was individually fitted to the harmonic regression model, minimizing the difference between the observed and estimated data. The projections were adjusted to local conditions using historical AT records, and the GCM AT was downscaled to the site-specific AT using the already constructed linear AT-AT relationship. The projected WTs under global warming scenarios A2 and B2, simulated by CCCma, were then presented as supporting data for future aquatic environment assessment. The primary conclusions are as follows:

- The averaged WT for the A2 scenario shows a ~3.0 °C increase for 2080-2010, whereas the WT for the B2 scenario shows a ~1.5 °C increase.
- The increases in the maximum WT are ~1.5 °C for the A2 scenario but ~1.0 °C for the B2 scenario for the years 2080-2100.

- The increases in the minimum WT are ~5 °C for the A2 scenario but ~3.0 °C for the B2 scenario for the years 2080-2100.
- The maximum/minimum WT shows a greater variation than the mean WT in the projection.
- The projected WT shows an increasing trend for the year 2100, but no seasonal patterns are found to increase.

This study will contribute to predictions of the impact of anthropogenic climate change on the water quality, ecosystems, and hydrologic regime in Korea. The methodology used for the study can be applied to any other location.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A4A01004846) for Khil-Ha Lee. This work was also supported by the KIOST Project (PE-98823) for Hong-Yeon Cho.

References

Alcamo, J., Doll, P., Kaspar, F., and Siebert, S. (1997). "Global change and global scenarios of water use and availability: an application of Water GAP 1.0." *Report A9701. Center for Environmental Systems Research*, University of Kassel, German

Barnett, V. and Lewis T. (1994). *Outliers in Statistical Data. Third Edition*, John Wiley & Sons.

Cho, H. and Lee, K. (2012). "Development of an air-water temperature relationship model to predict climate-induced future water temperature in estuaries" *Journal of Environmental Engineering-ASCE*, 138(5), 570-577.

Crisp, D.T., and Howson, G. (1982). "Effect of air temperature upon mean water temperature in streams in the north Pennines and English Lake District." *Freshwater Biol.*, 12, 359-367.

DeGaetano, A.T. (2001). "Spatial grouping of united states climate stations using a hybrid clustering approach." *International Journal of Climatology*, 2, 791-807.

Duan, Q.Y., Gupta, V.K., and Sorooshian, S. (1993). "Shuffled Complex Evolution approach for effective and efficient global minimization." *J. Optimiz. Theor. Appl.*, 76, 501–521.

Duan, Q.Y., Sorooshian, S., and Gupta, V.K. (1994). "Optimal use of the SCE-UA global optimization method for calibrating watershed models." *Journal of Hydrology*, 158, 265–284.

Hosking, J.R.M. and Wallis, J.R. (1997). *Regional frequency analysis-An approach based on L-moments*. Cambridge University Press, NY, USA, pp. 47

IPCC, 2007. *Climate change 2007: The physical science basis*. Cambridge University Press, Cambridge, UK

Johnson, F.A. (1971). "Stream temperatures in an alpine area" *Journal of Hydrology*, 14, 322-336.

Karl, T.R., Wang, W., Schleesinger, M.E., Knight, R.W., and Potman, D. (1990). "A Method of Relating General Circulation Model Simulated Climate to the Observed Local Climate. Part I: Seasonal Statistics" *Journal of Climate*, 3,1053-1079.

Little, R.J.A and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, Part I, 2nd Edition*, John-Wiley & Sons, Inc. NJ. USA

Mohseni, O., Ericson, T.R., and Stefan, H. (1999). "Sensitivity of stream temperature in the United States to air temperature projected under a global warming scenario" *Water Resources Research*, 35(12), 3723-3733.

Mohseni, O., Ericson, T.R., and Stefan, H. (2002). "Upper bounds for stream temperature in the contiguous United States."*Journal of Environmental Engineering*, 128(1), 4-11.

Mohseni, O., and Stefan, H. (1999). "Stream temperature/air temperature relationship: a physical interpretation" *Journal of Hydrology*, 218, 128-141.

Mohseni, O., Stefan, H., and Ericson TR (1998). "A nonlinear regression model for weekly stream temperature" *Water Resources Research*, 34(10), 2685-269.

Morril, J.C., Bales, R.C., and Conklin, M.H. (2005). "Estimating stream temperature from air temperature: Implications for future water quality" *Journal of Environmental Engineering*, 131(1), 139-146.

Nash, J.E. and Sutcliffe, J.V. (1970). "River flow forecasting through conceptual models, I-A Discussion of principles" *Journal of Hydrology*, 10, 282-290.

Nelder, J.A. and Mead, R.A. (1965). "A simplex method for function minimization" *Comput. J.*, 7, 308–313.

Oshima, N., Kato, H., and Kadokura, S. (2002). "An application of statistical downscaling to estimate surface air temperature in Japan" *Journal of Geophysical Research*, 107(14), 1-10

Pilgrim, J.M. and Stefan, H.G. (1995). "Correlation of Minnesota stream water temperatures with air temperatures" *Project. Rep. 382*. St Anthony Falls Lab., U of Minn., Minneapolis

School, J.T. and Pryor, S.C. (2001). "Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks" *International Journal of Climatology*, 21, 773-790.

Schubert, S. (1998). "Downscaling local extreme temperature changes in south-eastern Australia from the CSIRO Mark2 GCM" *International Journal of Climatology*, 18, 1419-1438.

Stefan, H.G. and Preud' home, E.B. (1993). "Stream temperature estimation from air temperature" *Water Resources Research*, 29(1), 27-45.

Stefan, H.G. and Sinokrot, B.A. (1993). "Projected global climate change impact on water temperatures in five north central US streams" *Climate change*, 24, 353-381.

Struyf, E., Damme, S.V., and Meire, P. (2004). "Possible effects of climate change on estuarine nutrient fluxes: a case study in the highly nutrified Schelde estuary (Belgium, The Netherland)" *Estuarine Coastal and Shelf Science*, 60, 649-661.

Webb, B.W. (1987). "The relationship between air and water temperatures for a Deven river" *Rep. Trans. Devaonshire Assoc. Adv. Sci.*, 119, 197-222.

Webb, B.W., and Nobilis, F. (1997). "Long term perspective on the nature of the air-water temperature relationship: A case study" *Hydrological Processes*, 11, 137-147.

Winkler, J.A., Palutik, J.P., Andersen, J.A., and Goodess, C.M. (1997) "The simulation of Daily Temperature Time Series from GCM Outtput. Part II: Sensitivity analysis of an Empirical Transfer function Methodology" *Journal of Climate*, 10, 2514-2532.

Tables

Table 1: Summary of WT data collected for the three studied locations

Site	Station	Lat	Lon	$T_w(^\circ C)$	$T_a(^\circ C)$	RH (%)	W_s (m/s)	Period
Ansan	S1	37.293	126.580	13.68	12.63	64.68	3.09	2002-2006
	S2	37.313	126.613	13.52	12.33	65.40	2.87	2002-2006
	S3	37.332	126.700	13.90	12.81	63.79	2.90	2002-2006
Masan	M1	35.196	128.575	16.48	15.78	63.81	2.72	2002-2006
	M2	35.201	128.578	16.49	15.29	65.80	2.88	2002-2006
Nakdong	N1	35.107	128.956	17.19	16.08	70.05	1.67	2004-2006

Table 2. Summary of the Korean meteorological stations

Station	Elevation (m)	Location	
		Lon. (degrees)	Lat. (degrees)
Daegwanrong	790.0	128° 43'	37° 40'
Seosan	25.2	126° 29'	36° 46'
Suwon	34.5	126° 59'	37° 16'
Cheonan	21.3	127° 07'	36° 46'
Chucheon	76.8	127° 44'	37° 54'
Gangneong	26.1	128° 53'	37° 45'
Seoul	85.5	126° 57'	37° 34'
Incheon	54.6	126° 37'	37° 28'
Wonju	150.7	127° 56'	37° 20'
Cheonju	56.4	127° 26'	36° 38'
Daejeon	62.6	127° 22'	36° 22'
Chupungrong	242.2	127° 59'	36° 13'
Andong	140.7	128° 42'	36° 34'
Pohang	1.3	129° 22'	36° 01'
Daegu	57.3	128° 37'	35° 53'
Jeonju	61.1	127° 09'	35° 49'
Gwangju	74.5	126° 53'	35° 10'
Busan	69.2	129° 01'	35° 06'
Jinju	27.1	128° 02'	35° 09'
Ganghwa	47.0	126° 27'	37° 42'
Mokpo	39.0	126° 23'	34° 49'
Cheongju	59.2	127° 26'	36° 38'
Jeju	22.6	126° 32'	33° 30'
Jejugosan	73.2	126° 10'	33° 17'
Hugsando	68.5	125° 45'	34° 68'

FIGURE CAPTIONS

Figure 1: Three study stations: Ansan, Masan, and Nakdong. All three sites are within river and ocean convergence areas.

Figure 2: Scatterplot of all station-averaged versus individual station data at the representative stations.

Figure 3: WT fitted to the harmonic function against the observed WT at the Masan station.

Figure 4: Statistical evaluation of the estimated WT against the observed WT. The figure shows the exceedance probability (EP) of the upper/lower quantiles of the estimated WT for a range of error norms.

Figure 5: Time series of the projected WT: (a) Masan, (c) Ansan, and (e) Nakdong for the A2 scenario and (b) Masan, (d) Ansan, and (f) Nakdong for the B2 scenario.

Figure 6: 20-year averaged values of the projected temperature for the A2 scenario at the representative stations: (a) mean AT, (b) mean WT, (c) maximum WT, and (d) minimum WT.

Figure 7: 20-year averaged values of the projected temperature for the B2 scenario at the representative stations: (a) mean AT, (b) mean WT, (c) maximum WT, and (d) minimum WT.

Figure 8: Annual variation of the projected values at the representative stations for the year 2100: (a) AT and (c) WT for the A2 scenario and (b) AT and (d) WT for the B2 scenario.

Figure 01: Three study stations: Ansan, Masan, and Nakdong. All
Click here to download high resolution image

Figure 02: Scatterplot of all station-averaged versus individual
Click here to download Figure: figure02.eps

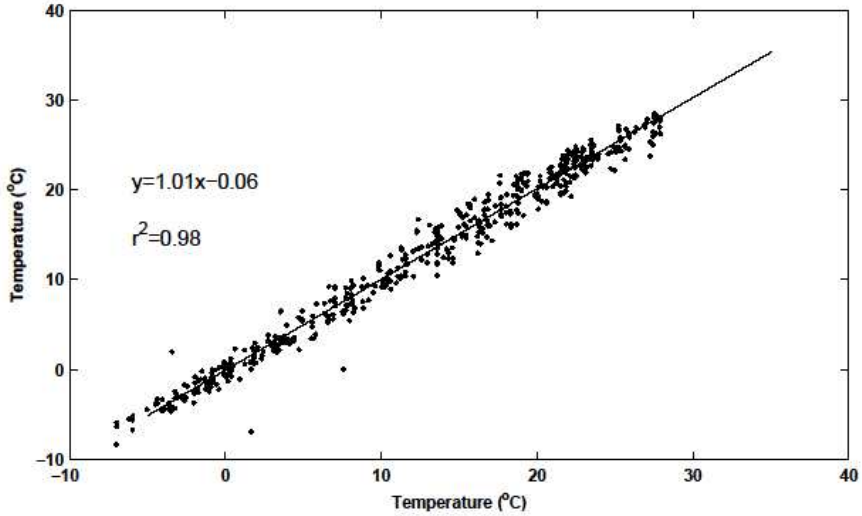


Figure 03: WT fitted to the harmonic function against the observ
Click here to download Figure: figure 03.eps

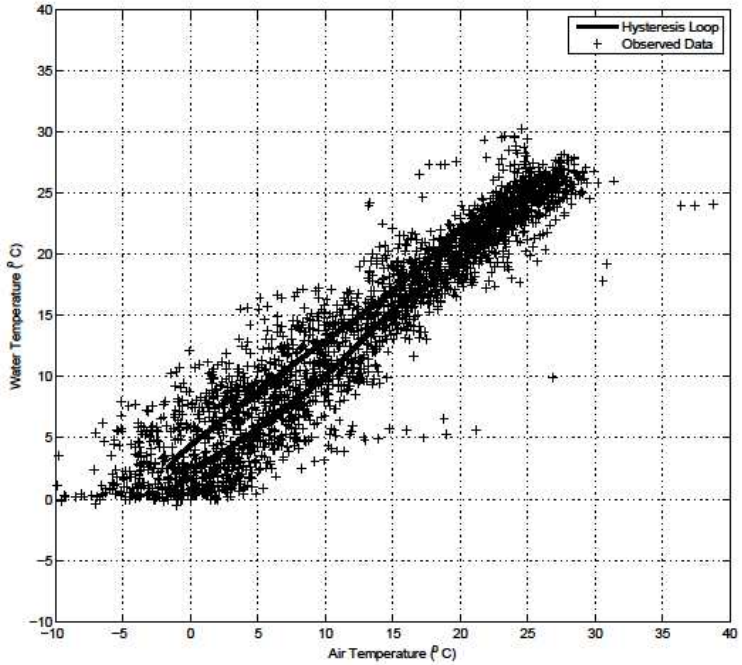


Figure 04: Statistical evaluation of the estimated WT against th
Click here to download Figure: figure 04.eps

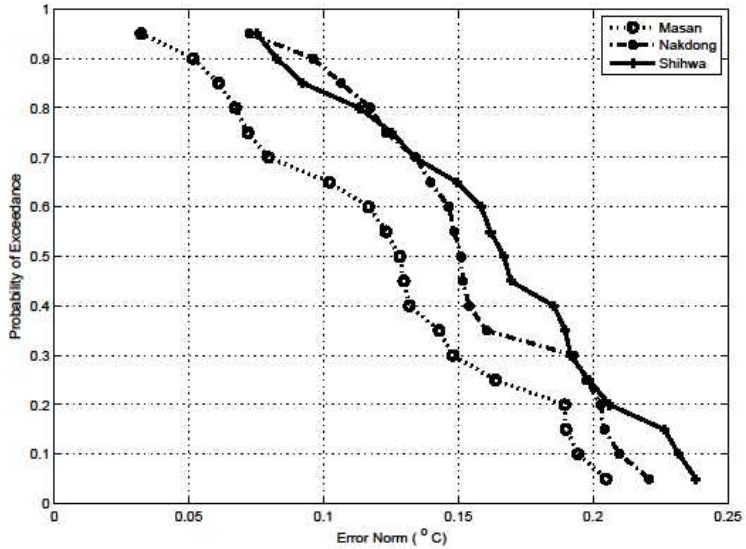


Figure 05: Time series of the projected WT: (a) Masan, (c) Ansan
Click here to download Figure: figure 05.eps

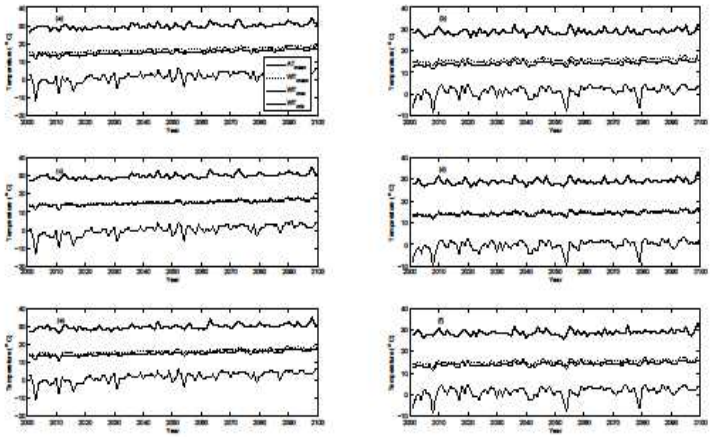


Figure 06: 20-year averaged values of the projected temperature
Click here to download Figure: figure 06.eps

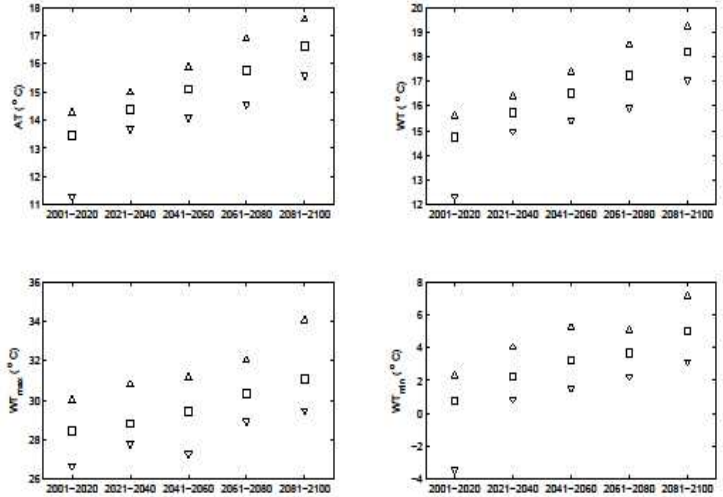


Figure 07: 20-year averaged values of the projected temperature
Click here to download Figure: figure 07.eps

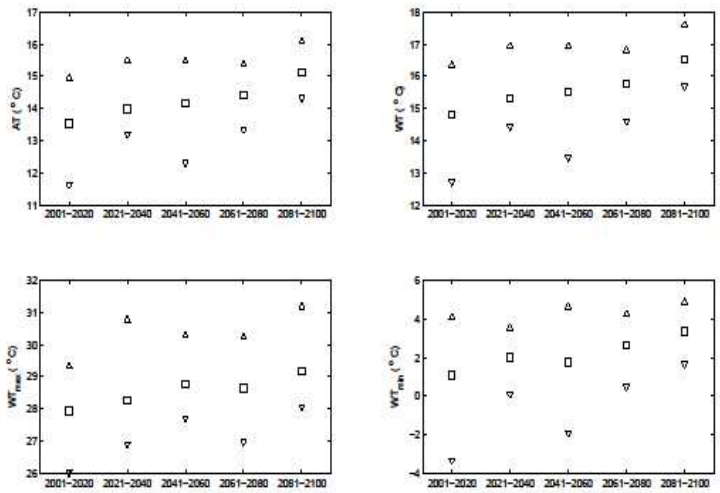
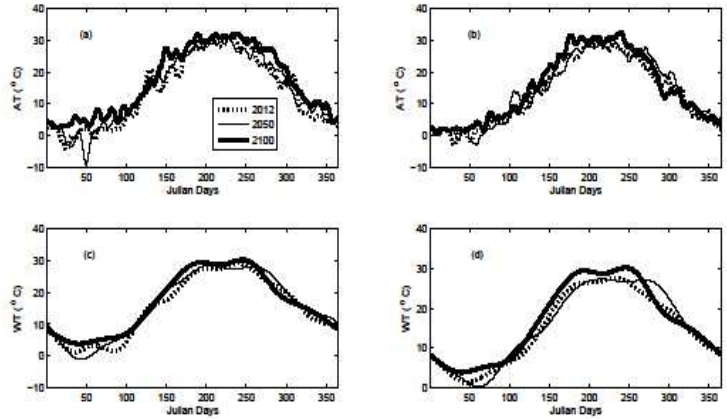


Figure 08: Annual variation of the projected values at the repre
Click here to download Figure: figure 08.eps



주 의

1. 이 보고서는 한국해양과학기술원에서 수행한 기본연구사업의 연구결과보고서입니다.

2. 이 보고서 내용을 발표할 때에는 반드시 한국해양과학기술원에서 수행한 기본연구사업의 연구결과임을 밝혀야 합니다.

3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 안 됩니다.