J. Soil Groundw. Environ. Vol. 21(6), p. 67~79, 2016 https://doi.org/10.7857/JSGE.2016.21.6.067
ISSN 1598-6438 (Print), ISSN 2287-8831 (Online)

≪ **Research Paper** ≫

가우시안 프로세스 회귀분석을 이용한 지하수 수질자료의 해석

구민호[1] · 박은규[2]* · 정진아[2] · 이헌민[2] · 김효건[3] · 권미진[4] · 김용성[5] · 남성우[5]
고준영[6] · 최정훈[7] · 김덕근[8] · 조시범[9]

[1]공주대학교 지질환경과학과
[2]경북대학교 지질학과
[3]벽산엔지니어링
[4]한국원자력환경공단
[5]지오그린21
[6]도화엔지니어링
[7]㈜ 지오이노베이션
[8]한국수자원공사
[9]한국농어촌공사

Applications of Gaussian Process Regression to Groundwater Quality Data

Min-Ho Koo[1] · Eungyu Park[2]* · Jina Jeong[2] · Heonmin Lee[2] · Hyo Geon Kim[3] · Mijin Kwon[4]
Yongsung Kim[5] · Sungwoo Nam[5] · Jun Young Ko[6] · Jung Hoon Choi[7] · Deog-Geun Kim[8] · Si-Beom Jo[9]

[1]Department of Geoenvironmental Sciences, Kongju National University, Kongju, Korea
[2]Department of Geology, Kyungpook National University, Daegu, Korea
[3]Byucksan Engineering, Seoul, Korea
[4]Korea Radioactive Waste Agency, Daejeon, Korea
[5]GeoGreen21 Co. Ltd., Seoul, Korea
[6]Dohwa Engineering, Seoul, Korea
[7]GeoInnovation, Daegu, Korea
[8]Korea Water Resources Corporation, Daejeon, Korea
[9]Korea Rural Community Corporation, Jeju, Korea

ABSTRACT

Gaussian process regression (GPR) is proposed as a tool of long-term groundwater quality predictions. The major advantage of GPR is that both prediction and the prediction related uncertainty are provided simultaneously. To demonstrate the applicability of the proposed tool, GPR and a conventional non-parametric trend analysis tool are comparatively applied to synthetic examples. From the application, it has been found that GPR shows better performance compared to the conventional method, especially when the groundwater quality data shows typical non-linear trend. The GPR model is further employed to the long-term groundwater quality predictions based on the data from two domestically operated groundwater monitoring stations. From the applications, it has been shown that the model can make reasonable predictions for the majority of the linear trend cases with a few exceptions of severely non-Gaussian data. Furthermore, for the data shows non-linear trend, GPR with mean of second order equation is successfully applied.

Key words : Groundwater quality, Trend analysis, Gaussian process regression, Theil-Sen estimator, Groundwater quality monitoring network

*Corresponding author : egpark@knu.ac.kr
Received : 2016. 9. 24 Reviewed : 2016. 11. 15 Accepted : 2016. 12. 21
Discussion until : 2017. 2. 28

1. 서 론

현재 국내에는 매년 많은 지하수 수질 및 수량자료가 누적되고 있는 상황이다. 환경부에서는 2014년 현재 지하수법 제 18조 및 지하수의 수질보전 등에 관한 규칙 제 9조에 기초하여 전국에 걸쳐 141개소 및 52개소의 전용측정망과 오염감시 전용측정망이 운영되고 있으며, 장기적으로 2030년까지 이들을 각각 1,305개 및 2,164개소로 확대하는 계획 하에 있다. 또한 국토교통부의 지하수관측망, 보조 지하수관측망, 해수침투 관측망, 지하수 관련 주요사업, 그리고 지자체의 지하수 시설로부터 얻어지는 지하수 수질자료까지 포함하면 그 양은 상당수에 이른다고 할 수 있다(김규범 외, 2010). 이러한 현황으로 볼 때, 지하수 수질자료를 체계적으로 관리하고 분석할 수 있는 다양한 방법론의 확보와 그 적용이 필요하다는 것은 주지의 사실이다.

지하수 수질의 장기적인 예측은 청정한 지하수질을 가지는 지역에 대해서는 지속적인 수질 유지방안 수립 그리고 오염이 인지되거나 오염이 우려되는 지역에 대해서는 중장기적인 개선대책을 수립할 수 있게 하는 등 지하수를 효율적으로 관리하는데 있어 근간이 된다고 할 수 있다. 또한 지하수 수질의 예측은 국가적 측면에서 지하수 수질을 관리하는 다양한 기반이라 할 수 있는 지하수 수질관측망의 설계, 관측망 설치, 관측자료의 수집, 관측 기술의 개선 등을 포함하여 가장 비용 효율성 및 활용성이 높다고 할 수 있다. 이러한 관점에서 장기적 추이에 대한 예측 도구의 개발은 지속적으로 이루어져야 한다.

국내 대부분의 지하수 수질 추세 분석 및 예측 방법론은 비모수적(non-parametric) 방법으로 분류되는 Mann-Kendall 경향성 분석 및 Theil-Sen 기울기 추정을 통하여 이루어져 왔으며 이러한 적용 역시도 그 사례가 매우 드문 상황이다. 김규범 외(2010)는 Theil-Sen 기울기 추정을 통하여 토지이용 별 일반오염물질과 전기전도도 추세를 10여년의 기간에 대하여 분석하고 수질 개선 및 수질 악화를 분류한 바 있다. 한국환경공단과 환경부는 2012년 국가 지하수수질전용측정망 지하수 운영관리 보고서(2012) 중 수질관리 프로세스 개발의 일환으로 Mann-Kendall 경향성 분석 및 Theil-Sen 추정을 이용하여 수질측정망 91개소 240 지점에 대하여 11개 항목(총 2,640개)에 대한 분석을 실시한 바 있다. 이들 사례를 제외하고는 지하수 수질에 대한 경향성 분석은 매우 드문 상황이며, 적용된 방법론 역시 매우 제한적이다.

지하수 수질의 추세를 분석하는 비모수적 기법의 대안으로 모수적(parametric) 기법을 고려할 수 있다. 기존 연구들에서 비모수적 방법의 선호 이유는 모수적 방법의 근간이 되는 선형 회귀분석의 주요 가정인 잔차의 정규성(Gaussianity)이 성립하지 않는 경우가 일반적이기 때문이다(Helsel and Hirsch, 1988). 그러나 모수적 방법이 갖는 비선형 추세 활용 및 주기적 변동성 표현 가능성 등 예측에의 유연성과 외부적 요인이 예측에 반영될 수 있다는 점 등은 여전히 모수적 방법의 강점이라 할 수 있다. 또한 모수적 방법은 자료의 형태 및 수에 상관없이 예측을 수행할 수 있어, 상대적으로 많은 자료가 요구되는 비모수적 방법의 대안이 될 수 있다.

본 연구에서는 지하수 장기예측을 위한 모수적 기법의 일환으로 가우시안 프로세스 회귀분석(Gaussian Process Regression, 이하 GPR)을 활용하였다. GPR 기법은 예측과 동시에 예측에 따른 불확실성을 제공할 수 있는 기법으로 지하수 수질예측의 신뢰성을 제고할 수 있다는 측면에서 향후 활용성이 큰 기법으로 판단된다. 본 연구에서는 전국 지하수 수질에 걸친 대대적인 기법의 적용을 목표로 하지는 않았다. 다만, 가상 및 실제 자료를 이용한 GPR 기법의 적용 사례 소개를 통하여 기법의 특성 및 향후 지하수 수질 예측에 있어 GPR 기법의 적용 가능성을 보이는 것을 목표로 설정하였다. 이를 위하여 본 연구에서는 기존 모수적을 비모수적 기법과 비교함으로써 한계성 및 유용성을 우선 설명하고, GPR 기법의 이론을 간략하게 소개하며, 가상의 자료를 통하여 기존 모수적 기법과의 비교를 수행하고, 마지막으로 실제 자료에의 적용을 통해 GPR 기법의 적용성 및 예측 유연성을 보여 다양한 시사점을 도출하고자 한다.

2. 일반 선형 회귀분석 및 수질자료 분석의 적용성과 한계성

선형 회귀분석은 종속변수(dependent variable)인 지하수 특정 수질농도와 지하수질 농도를 결정하는 다양한 인자인 설명변수(explanatory variable) 간의 관계를 설명하는 통계적 기법의 일종이다. 이러한 회귀분석 중 설명변수가 하나인 것을 단순회귀분석이라 하며, 하나 이상의 설명변수가 존재할 경우 이를 다중회귀분석이라 부른다. 선형회귀분석은 단순히 설명변수와 종속변수 간 직선의 관계를 의미하는 것은 아니며, 설명변수들의 비선형 변환을 통하여 어떠한 형태의 커브도 표현이 가능하다.

주어진 자료가 $D = [(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)]$ 와 같을 때, 자료를 설명하는 모델이

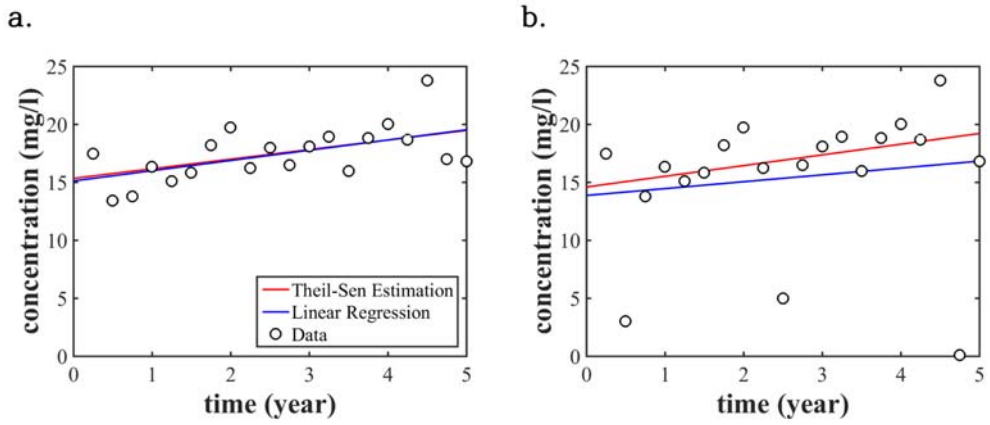

Fig. 1. Comparative applications of the parametric (linear regression, blue line) and non-parametric method (Theil-Sen estimator, red line) for the groundwater quality data showing (a) Gaussian and (b) non-Gaussian distribution of the residuals

$$h_\theta(\mathbf{x}) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_d x_d = \theta^{\mathrm{T}}\mathbf{x} \tag{1}$$

와 같이 선형식으로 표현할 수 있다고 가정할 수 있으며, 여기서 θ는 가중치로 이루어진 벡터이며 $\mathbf{y}$는 관찰된 자료 벡터 그리고 $\mathbf{x}$는 설명변수 벡터를 의미한다. 선형식을 이용한 최대가능도 추정량(MLE, Maximum Likelihood Estimation)을 얻기 위해서는 일반적으로 정규분포 함수를 사용하며 다음의 식과 같이 쓸 수 있다.

$$p(\mathrm{D}|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_i - \theta^{\mathrm{T}}\mathbf{x}_i)^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{\sum_{i=1}^{n}(y_i - \theta^{\mathrm{T}}\mathbf{x}_i)^2}{2\sigma^2}\right] \tag{2}$$

위의 식에서 주어진 조건부 확률은 식 (2)의 음-대수-우도함수(negative log-likelihood)로부터

$$J(\theta) = \sum_{i=1}^{n}(y_i - \theta^{\mathrm{T}}\mathbf{x}_i)^2 \tag{3}$$

와 같이 목적함수가 얻어진다. 따라서, 식 (3)이 최소일 경우 식 (2)는 최대가 된다. 일반적으로 위의 식을 만족하는 θ를 얻기 위한 방법에는 정규방정식(normal equation)을 이용하는 방법과 기울기 하강 기법(gradient descent method)을 이용하는 방법이 있다. 이 중 정규방정식을 이용한 방법은

$$J(\theta) = \sum_{i=1}^{n}(y_i - \theta^{\mathrm{T}}\mathbf{x}_i)^2 = (\mathbf{y} - \theta^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \theta^{\mathrm{T}}\mathbf{x})$$

$$= (\mathbf{y} - \mathbf{A}\theta)^{\mathrm{T}}(\mathbf{y} - \mathbf{A}\theta) = \|\mathbf{y} - \mathbf{A}\theta\|^2 \tag{4}$$

와 같이 접근하는 방법이며, 여기서 행렬 $\mathbf{A}$를 설계행렬(design matrix)이라 부른다. 위의 목적함수인 식 (3)이 최소화되기 위해서는

$$\nabla_\theta \|\mathbf{y} - \mathbf{A}\theta\|^2 = -2\mathrm{A}^{\mathrm{T}}\mathrm{y} + 2\mathrm{A}^{\mathrm{T}}\mathbf{A}\theta = 0 \tag{5}$$

의 조건을 만족하여야 하며 결과적으로

$$\theta_{\mathrm{MLE}} = (\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{T}}\mathrm{y} \tag{6}$$

을 얻게 된다.

그러나 일반 선형 회귀분석은 잔차의 정규분포를 가정하며, 이를 충족하지 않았을 경우 예측의 제한성(i.e., 예측 결과가 최대가능도 추정량이 아님)이 따르게 된다. 또한 일반 선형 회귀분석은 이상 농도(outlier)에 일반적으로 매우 민감하다는 한계를 지니고 있다. 이를 설명하기 위하여 선형 회귀분석에 의한 예측과 비모수적 방법인 Theil-Sen 추정을 비교하였다. 다음의 예는 선형회귀분석의 한계성을 보여준다. Fig. 1은 시간에 따른 가상의 수질 자료 변화를 보여준다. 관측은 1년에 4회 이루어진다. 가상의 지하수 수질자료를 생성하기 위하여 식 $C = 15 + 0.2 \times t_{1/4} + Z\ (Z \sim N(0, 3^2))$가 이용되었다. 해당 식은 관측 직전의 농도가 15 mg/l이고 분기($t_{1/4}$) 당 평균 0.2 mg/l의 상승추세를 보이는 오염원의 농도를 의미하며, 평균이 0 및 표준편차가 3 mg/l의 정규 분포를 따르는 잔차($N(0, 3^2)$)가 포함됨을 가정하였다(Fig. 1a). 또한 잔차가 비정규분포를 따르는 자료를 만들기 위하여, 상기의 자료에서 관측 1년차 2번째 분기의 농도는 3 mg/l, 관측 3년차 2번째 분기의 농도는 5 mg/l, 그리고 5년차 3번째 분기의 농도는

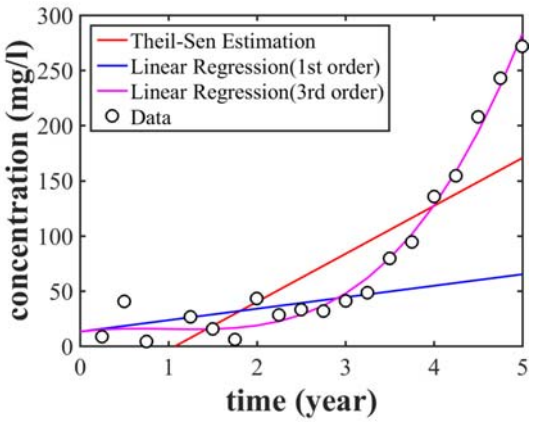

Fig. 2. Comparative applications of Theil-Sen estimator (red line), 1st order linear regression (blue line), and 3rd order linear regression (magenta line) to non-linear trend of the concentration data.

0.1 mg/l로 설정하였다(Fig. 1b).

Fig. 1에서 보는 바와 같이 잔차가 정상분포를 보이는 Fig. 1a에 대해서는 모수적 기법인 선형 회귀분석에 의한 예측과 비모수적 기법인 Theil-Sen 추정에 의한 예측이 유사함을 확인할 수 있다. 예측된 분기 당 농도 변화율인 추세선의 기울기 역시 선형 회귀분석에서는 0.22 mg/l 및 Theil-Sen 추정에서는 0.21 mg/l로 예측되어 가정된 실제 값인 0.2 mg/l에 가까운 유사한 값이 산정되었다. 그러나 잔차가 비정상분포를 보이는 Fig. 1b의 경우 선형 회귀분석에 의한 예측과 Theil-Sen 추정에 의한 예측은 뚜렷하게 서로 다르다. 또한 기울기의 경우 선형 회귀분석에서는 0.15 mg/l 및 Theil-Sen 추정에서는 0.23 mg/l로 예측되어 Theil-Sen 추정이 가정된 실제 값인 0.2 mg/l에 더 가까운 것으로 분석되었다. 이러한 분석 결과는 Theil-Sen 추정이 이상 농도에 대한 저항력이 더 크며 안정적으로 실제 프로세스에 해당하는 예측을 보다 유사하게 하는 반면 선형 회귀분석에 의한 예측은 이상 농도에 매우 민감하게 예측한다는 것을 설명한다.

Fig. 2는 수질자료가 비선형적 추세를 뚜렷하게 보일 경우 Theil-Sen 추정, 1차 및 3차 선형회귀 분석을 각각 적용한 결과이다. 단, 가상 자료의 생성을 위하여 이용된 3차 선형회귀 분석은 다음의 식을 이용하였다.

$$C = \theta_0 + \theta_1 t + \theta_2 t^2 + \theta_3 t^3 \tag{7}$$

Fig. 2는 Theil-Sen 추정 및 1차 선형회귀 분석은 비록 농도의 상향 추세를 예측하나 효과적인 예측에 실패한 반면, 3차 선형회귀 분석은 농도의 추세를 잘 예측하고 있음을 보여준다. 실제로 많은 자연적 및 인위적 영향 하에 놓인 지하수의 장기적 수질농도 변화는 비선형 추세를 보이는 것이 보다 일반적이며, 위의 분석 결과는 이에 대하여 선형적 예측(Theil-Sen 추정 및 1차 선형회귀)을 통해 수질을 예측할 경우 그 결과가 실제 수질변화를 반영하지 않을 가능성이 매우 큼을 보여준다.

결론적으로 자료가 선형의 추세를 보일 경우 Theil-Sen 추정을 하는 것이 보다 바람직하며 이상치를 보이지 않을 경우 선형 회귀분석이 이루어질 수 있다. 그러나 자료의 추세가 뚜렷하게 비선형적일 경우, Theil-Sen 추정 혹은 1차 선형회귀 분석은 적용이 불가하며 보다 높은 차수를 갖는 선형회귀 분석을 실시하는 것이 바람직하다.

3. GPR 기법을 이용한 지하수 수질자료의 분석

커널기법(kernel method)에 기초한 기계학습 기법의 일종인 GPR 기법(Rasmussen and Williams, 2006)은 종속변수인 지하수 수질농도가 가우시안 프로세스를 따를 경우 실시하는 회귀분석 기법의 일종이며 많은 선진 기계학습기법들(i.e. 로지스틱 회귀분석, 인공신경망, 지지기반벡터 등)과 유사한 구조적 특성을 갖는다. 보다 구체적으로, GPR 기법을 이용한 수질자료 분석은 지하수 내 특정 항목 농도(C^*)가 오차를 포함하는 확률변수(i.e. $C^* = C + \varepsilon$)라고 가정하였을 경우, 오차를 제거한 기대 농도는 평균과 오차 간의 공분산 함수로 표현될 수 있으며, 다시 이러한 오차 공분산은 커널함수로 해석될 수 있다고 가정할 때 수질에 대한 예측을 수행하는 베이지안(Bayesian) 추론 모델이다. GPR을 지하수 수질자료 분석에 적용할 경우 장기적인 예측뿐만 아니라 예측의 질 또는 예측의 불확실성을 동시에 얻을 수 있어, 보다 심도 있는 수질 예측 결과의 활용이 가능하며 많은 발전 여지를 지니고 있다(Rasmussen, 2004). 수질예측 분야에서는 근래에 들어 일부 적용이 이루어지고 있어 아직까지 많은 활용 사례는 없는 상황이며, 하천유량의 예측(Sun et al., 2014), 하천 녹조류 변화 예측(Bazi et al., 2012), 하천 온도변화 예측(Grbiet al., 2013) 등에 이용되기 시작하고 있다.

GPR의 이론적 기반은 다음과 같다. 시간벡터 t_o 동안 관측된 특정 농도 시계열 자료를 $\mathbf{X}_o$ 그리고 시간벡터 동안 t_p 예측하여야 할 특정 농도 시계열 자료를 $\mathbf{X}_p$라 하였을 때 이 자료 벡터를 종합하여 $\mathbf{X}$라 표현할 수 있으며 각 시간에서의 기댓값인 μ와 함께 각각 다음과 같이 쓸 수 있다.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_o \\ \mathbf{X}_p \end{bmatrix} \quad \text{및} \quad \mu = \begin{bmatrix} \mu_o \\ \mu_p \end{bmatrix} \tag{8}$$

여기서 자료의 공분산 $\mathbf{C}$는 $(\mathbf{X}-\mu)^T(\mathbf{X}-\mu)$ 이며 따라서 관측된 농도 시계열 자료(기호 o) 및 예측하여야 할 특정 농도 시계열 자료(기호 p)에 대해 분할된 분산 행렬은

$$\mathbf{C}=\begin{bmatrix}\mathbf{C}_{oo}\,\mathbf{C}_{op}\\ \mathbf{C}_{po}\,\mathbf{C}_{pp}\end{bmatrix} \tag{9}$$

와 같이 쓸 수 있다. 이 공분산은 다시 LDU 분해법에 의해

$$\mathbf{C}=\begin{bmatrix}\mathbf{C}_{oo}\,\mathbf{C}_{op}\\ \mathbf{C}_{po}\,\mathbf{C}_{pp}\end{bmatrix}=\begin{bmatrix}\mathbf{I}\,\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\\ \mathbf{0}\quad\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{oo}-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{C}_{po} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{pp}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{C}_{pp}^{-1}\mathbf{C}_{po}\,\mathbf{I}\end{bmatrix} \tag{10}$$

와 같이 분해할 수 있으며, 위의 식 (9)는 다시 Schur 여수(complement) 정리를 이용하여

$$\mathbf{C}^{-1}=\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{C}_{pp}^{-1}\mathbf{C}_{po}\,\mathbf{I}\end{bmatrix}\begin{bmatrix}(\mathbf{C}_{oo}-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{C}_{po})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{pp}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I}\,\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\\ \mathbf{0}\quad\mathbf{I}\end{bmatrix} \tag{11}$$

와 같이 쓸 수 있다. 여기서 $\mathbf{Y}_o=\mathbf{X}_o-\mu_o$ 그리고 $\mathbf{Y}_p=\mathbf{X}_p-\mu_p$라 정의하고 각각이 정규분포를 따른다고 가정하면, 동시 확률분포는

$$p(\mathbf{Y}_o,\mathbf{Y}_p)\propto\exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{Y}_o\\ \mathbf{Y}_p\end{bmatrix}^T\mathbf{C}^{-1}\begin{bmatrix}\mathbf{Y}_o\\ \mathbf{Y}_p\end{bmatrix}\right) \tag{12}$$

와 같이 주어지며, 위의 식 (11)을 식 (12)에 대입하여 전개하면

$$p(\mathbf{Y}_o,\mathbf{Y}_p)\propto$$

$$\exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{Y}_o-\mathbf{Y}_p\mathbf{C}_{pp}^{-1}\mathbf{C}_{po}\\ \mathbf{Y}_p\end{bmatrix}^T\begin{bmatrix}(\mathbf{C}_{oo}-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{C}_{po})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{pp}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_o-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{Y}_p\\ \mathbf{Y}_p\end{bmatrix}\right)$$

이며, 이는 다시

$$p(\mathbf{Y}_o,\mathbf{Y}_p)\propto$$

$$\exp\left[-\tfrac{1}{2}(\mathbf{Y}_o-\mathbf{Y}_p\mathbf{C}_{pp}^{-1}\mathbf{C}_{po})^T(\mathbf{C}_{oo}-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{C}_{po})^{-1}(\mathbf{Y}_o-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{Y}_p)\right]$$

$$\times\exp\left(-\frac{1}{2}\mathbf{Y}_p^T\mathbf{C}_{pp}^{-1}\mathbf{Y}_p\right) \tag{13}$$

가 된다. 따라서 Bayes의 정리 $p(\mathbf{Y}_a,\mathbf{Y}_b)=p(\mathbf{Y}_a|\mathbf{Y}_b)\,p(\mathbf{Y}_b)$ 에 의해

$$p(\mathbf{Y}_o,\mathbf{Y}_p)\propto$$

$$\exp\left[-\tfrac{1}{2}(\mathbf{Y}_o-\mathbf{Y}_p\mathbf{C}_{pp}^{-1}\mathbf{C}_{po})^T(\mathbf{C}_{oo}-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{C}_{po})^{-1}(\mathbf{Y}_o-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{Y}_p)\right] \tag{14}$$

와 같이 쓸 수 있다. 위의 식 (14)은 정규분포 함수의 형태를 가지며 이로부터 다변량 가우시안 정리(multivariate Gaussian theorem)를 통해 $p(\mathbf{Y}_o|\mathbf{Y}_p)=N(\mathbf{m},\mathbf{D})$ 의 평균 및 공분산은 각각

$$\mathbf{m}=\mu_o-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}(\mathbf{X}_p-\mu_p) \tag{15-1}$$

$$\mathbf{D}=\mathbf{C}_{oo}-\mathbf{C}_{op}\mathbf{C}_{pp}^{-1}\mathbf{C}_{po} \tag{15-2}$$

와 같이 주어진다. 관찰에 의해 수정된 평균 $\mathbf{m}$ 및 분산 $\mathbf{D}$는 공분산을 모델화하는 커널함수 및 초기 예측으로부터 반복적 연산에 의해 얻을 수 있다. 여기서, 커널함수는 상호 떨어져 있는 자료 간의 유사성을 설명하는 함수로 t_1과 t_2에서 관찰된 두 자료인 Z_1과 Z_2의 공분산은

$$\mathrm{Cov}(Z_1,Z_2)=k(t_1,t_2)=k(|t_1-t_2|) \tag{16}$$

와 같이 t_1과 t_2 간의 거리를 유일한 변수로 하는 함수에 의해 설명할 수 있다고 가정한다. 일반적으로 커널 함수에는 방사상 기반함수(radial basis function, RBF)가 이용되며, 본 연구에서 이용된 RBF는 매우 일반적인 형태로

$$k(t_1,t_2)=\alpha\exp\left[-\frac{(t_1-t_2)}{2\beta^2}\right] \tag{17}$$

와 같다. 또한 RBF 모델 파라미터인 α와 β는 자료를 통하여 경험적으로 예측되어야 하는 상수로, 일반적으로 자료마다 특성적인 값의 범위를 가지고 있으며 최적화 되어야 한다. 여기서 β의 경우 서로 다른 기간에 측정한 수질이 상호 연계성을 지니기 위한 시간적 거리를 의미하는 상관거리(correlation scale)이며, 따라서 β보다 큰 시간적 차이를 두고 측정한 두 수질관측 사이에는 상관성이 없다고 할 수 있다.

만약, 평균값에 선형 또는 비선형의 추세가 있을 경우 GPR 기법은 다소간의 이론적 수정이 필요하며 이의 내용은 Rasmussen(2006) 및 Murphy(2012)에 매우 상세히 소개되어 있다. 따라서, 본 연구에서는 선형 및 비선형 추세를 예측 평균으로 한 GPR 기법 이론의 설명은 생략하기로 한다.

본 연구의 분석을 위하여 Theil-Sen 추정, 선형회귀분석, 및 GPR의 전산코드가 MATLAB 8.6(R2015b)을 기반으로 제작되었으며, 본 논문의 주 저자와의 협의를 통하여 제한적으로 제공 가능하다.

GPR의 예측 능력 및 지하수 수질자료 적용성을 살펴보기 위하여 가상의 수질자료를 생성하고 이에 대한 적용을 실시하였다. 가상 수질자료는 선형 추세를 보이는 자료와

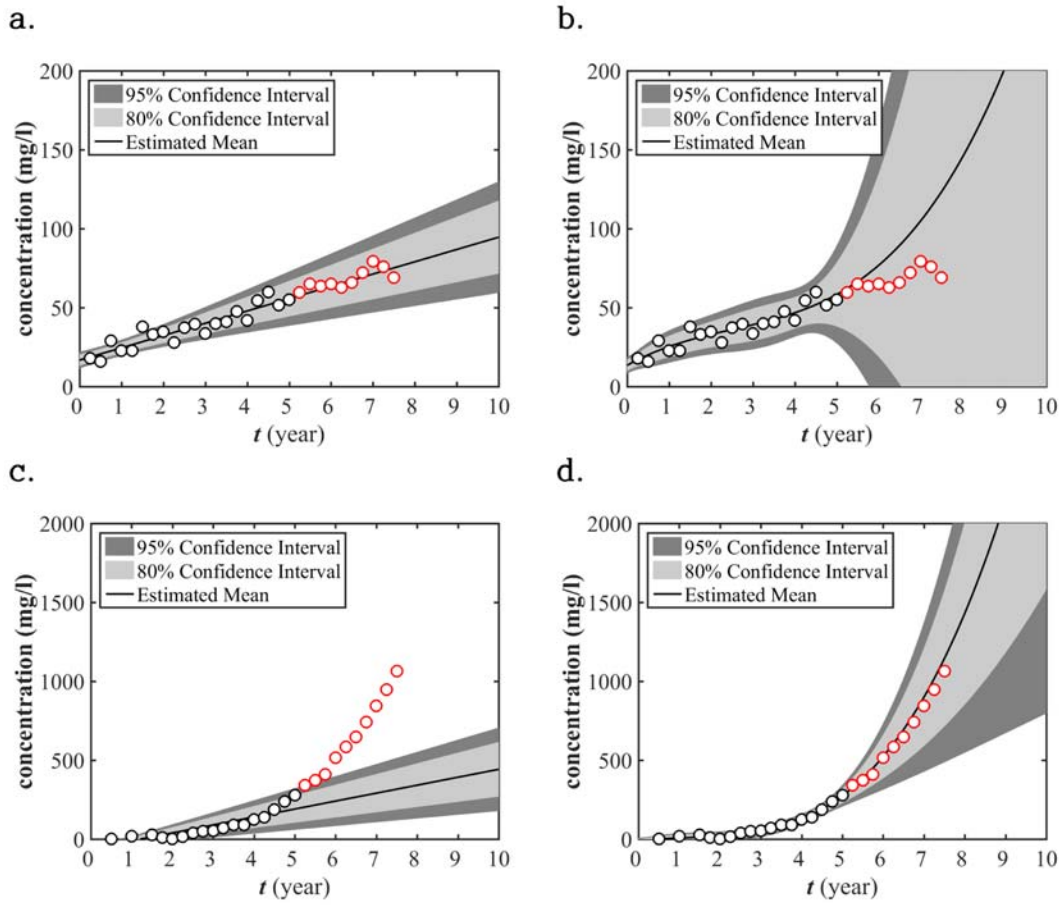

Fig. 3. Validation of the prediction accuracy of GPR method based on the hypothetical groundwater quality data (black circle-training dataset; red circle-test dataset) where black lines indicate the estimated mean trends and dark- and light-gray parcels indicate the 95% and 80% confidence intervals, respectively. Each panel show the result applying (a) 1st order linear regression, (b) 2nd order linear regression to 1st order linear trend of the data, and (c) 1st order linear regression, (d) 2nd order linear regression to 2nd order linear trend of the data.

비선형 추세를 보이는 자료 두 종류로 아래의 식을 이용하여 각각 제작하였다.

$$C = 15 + 2 \times t_{1/4} + Z,\ Z \sim N(0, 5^2)\ \text{(linear trend)} \quad (18\text{-}1)$$

$$C = 5 + 0.01 \times t_{1/4}^{3.4} + Z,\ Z \sim N(0, 10^2)\ \text{(non-linear trend)} \quad (18\text{-}2)$$

농도 자료는 관측 시작부터 총 40분기(10년)에 대하여 생성하였으며 이 중 앞선 20분기(5년)의 자료를 훈련자료(검은색 원)로 활용하였고, 이후 10분기(2.5년)에 걸친 농도를 예측결과에 대한 검증자료(빨간색 원)로 이용한 후, 나머지 10분기를 예측기간으로 설정하였다.

첫 번째 경우는 가상의 지하수 수질자료가 선형을 따르며 수질자료에 이상 농도가 없는 상황이다(식 (18-1)). 지하수 수질 예측을 위하여 GPR의 예측평균이 각각 1차식 ($\overline{C} = \theta_0 + \theta_1 t$) 및 3차식 ($\overline{C} = \theta_0 + \theta_1 t + \theta_2 t^2 + \theta_3 t^3$)을 따름을 가정한 후 예측이 수행되었으며 그 결과는 각각 Fig. 3a-b와 같다. 예측을 위해 가정된 파라미터 α 및 β (식 (17))는 각각 0.1 및 0.5이다. Fig. 3a와 같이 선형의 수질추세에 대하여 1차식을 GPR의 예측평균으로 사용한 경우 검증자료는 모두 예측된 평균 추세선 상 또는 인근에 위치하는 것을 확인할 수 있다. 또한 80% 및 95%로 표현한 신뢰구간(confidence interval)과 검증자료를 비교하여 보았을 때에도 모든 검증자료가 80% 내에 존재하고 신뢰구간의 크기도 상대적으로 작아 좋은 예측이 이루어졌다고 할 수 있다. Fig. 3b의 경우 3차식을 예측평균으로 사용한 GPR 결과이며 검증자료와 예측평균선의 차이가 큰 것으로 나타났다. 또한, 비록 검증자료가 80% 신뢰구간 내에 모두 위치하나 신뢰구간의 폭이 커 좋은 예측이 이루어졌다고 판단하기 어렵다. Fig. 3c-d는 가상의 지하수 수질자료가 비선형적 추이를 지니는 경우이다(식 (18-2)). 마찬가지로 GPR 예측평균에는 1차식(Fig. 3c) 및 3차식(Fig. 3d)이 각각 이용되었다. 1차식을 예측평균으로

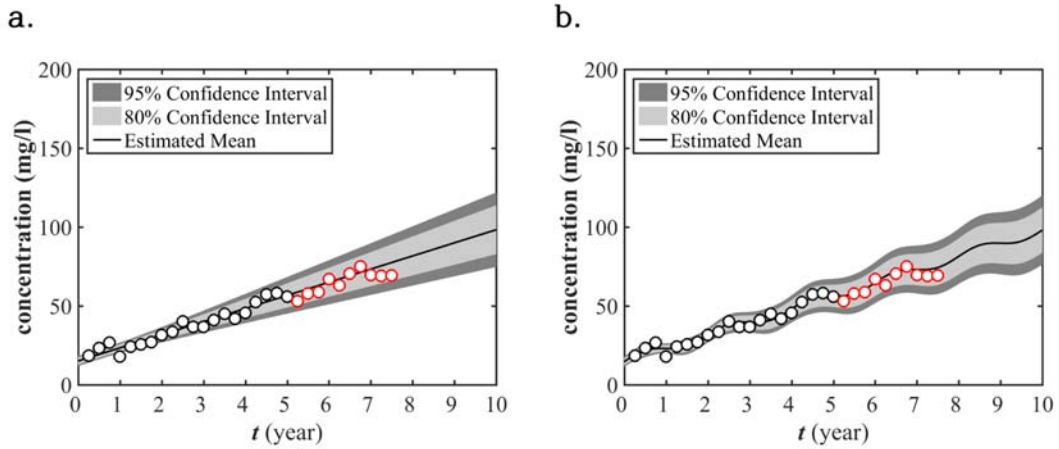

Fig. 4. Prediction accuracy of GPR method on the seasonally variable hypothetical data based on linear regression with: (a) 1st order linear function and (b) periodic function. Black circles, red circles, black lines, dark- and light-gray parcels indicate training dataset, test dataset, the estimated mean trends and the 95% and 80% confidence intervals, respectively.

한 GPR 예측의 경우 대부분의 검증자료가 95% 신뢰구간 바깥에 놓이게 되어 잘못된 예측이 이루어졌음을 쉽게 판단할 수 있다. 그러나 3차식을 예측평균으로 이용하였을 경우 검증자료와 예측된 평균 추세선이 매우 가까우며 모두 80% 신뢰구간 내에 놓이게 되어 양호한 예측이 이루어진 것으로 판단할 수 있다.

이러한 가상 자료를 통한 GPR 예비 적용을 통하여 추세에 대한 선형 내지 비선형 파악은 매우 중요하다는 사실을 알 수 있다. 또한, 수질자료를 설명하는 예측평균 식을 선정함에 있어서, 높은 차수의 식을 이용할수록 보다 유연하게 자료를 예측을 하는 측면이 있는 반면 이에 따른 예측 불확실성 역시 크게 증가할 수 있으므로 가급적 자료를 잘 설명할 수 있는 낮은 차수의 식을 이용한 예측이 필요하다.

Fig. 4는 주기성을 포함하는 가상의 지하수 수질을 GPR로 예측한 결과를 보여준다. 가상의 지하수 수질에 주기성을 부여하기 위하여 아래와 같은 식이 이용되었다.

$$C = 15 + 2 \times t_{1/4} + 3 \times \sin\left(\frac{\pi t_{1/4}}{4}\right) + Z,\ Z \sim N(0, 5^2) \qquad (19)$$

만약 이러한 주기성이 계절적 변동이나 인근 지하수 이용 양상 등과 같이 충분히 예측할 수 있을 경우, 혹은 자료를 통하여 이러한 주기성이 분석될 수 있을 경우 이를 GPR의 예측평균에 반영할 수 있다. 아래의 Fig. 4a는 이러한 수질자료의 주기성을 고려하지 않은 경우의 분석결과이며, Fig. 4b는 주기성을 고려한 예측 결과이다. 그림을 통해 확인할 수 있는 바와 같이 주기성을 반영한 예측에서 검증자료와 예측된 평균선 간의 이격이 더 작게 나타나며, 신뢰구간 역시 계절적 영향이 반영되어 보다 정확한 예측이 되었다고 할 수 있다.

4. 실제 지하수 수질자료 적용

GPR 기법의 실제 지하수 수질자료 적용을 위하여 2007년에 설치되어 2008년부터 관측이 시작된 총 2개소 지하수 수질전용 측정망(강릉장현 및 평창유천) 자료 중 지표 유입 수질에 가장 큰 영향을 받을 것으로 판단되는 최상부 심도에 대한 분석을 수행하였다. 분석에 사용된 기간은 모든 관측소에서 동일하게 2008년부터 2014년 까지 총 7년 동안이며 연간 4회 측정을 이용하였다. 이 중 GPR의 훈련자료로는 6년(24분기) 동안의 자료가 활용되었으며 나머지 1년(4분기) 동안의 자료는 예측에 대한 검증 자료로 이용되었다. 모든 분석 및 예측은 1차식을 예측평균으로 이용한 GPR을 기본으로 수행되었으며 일부 비선형적 변화추세가 뚜렷한 자료에 대하여 2차식 예측평균이 이용되었다.

관측된 지하수 수질자료의 활용 시 미 관측 자료는 훈련에서 제외하였으며 관측이 이루어졌으나 불검출된 경우 0 mg/l의 농도를 부여하였다. 분석 대상 관측 항목으로는 지하수위, pH, TDS, T, 염소이온, 황산이온, 질산성질소, 암모니아성질소이다.

4.1. 강릉장현 관측소

강릉장현 관측소는 강원도 강릉시 장현동 74-7 강원도 가축위생사업소 내에 위치하고 있으며 측정심도의 매질은 편마암으로 이루어져 있다. 관측소의 최상부 관측 심도는

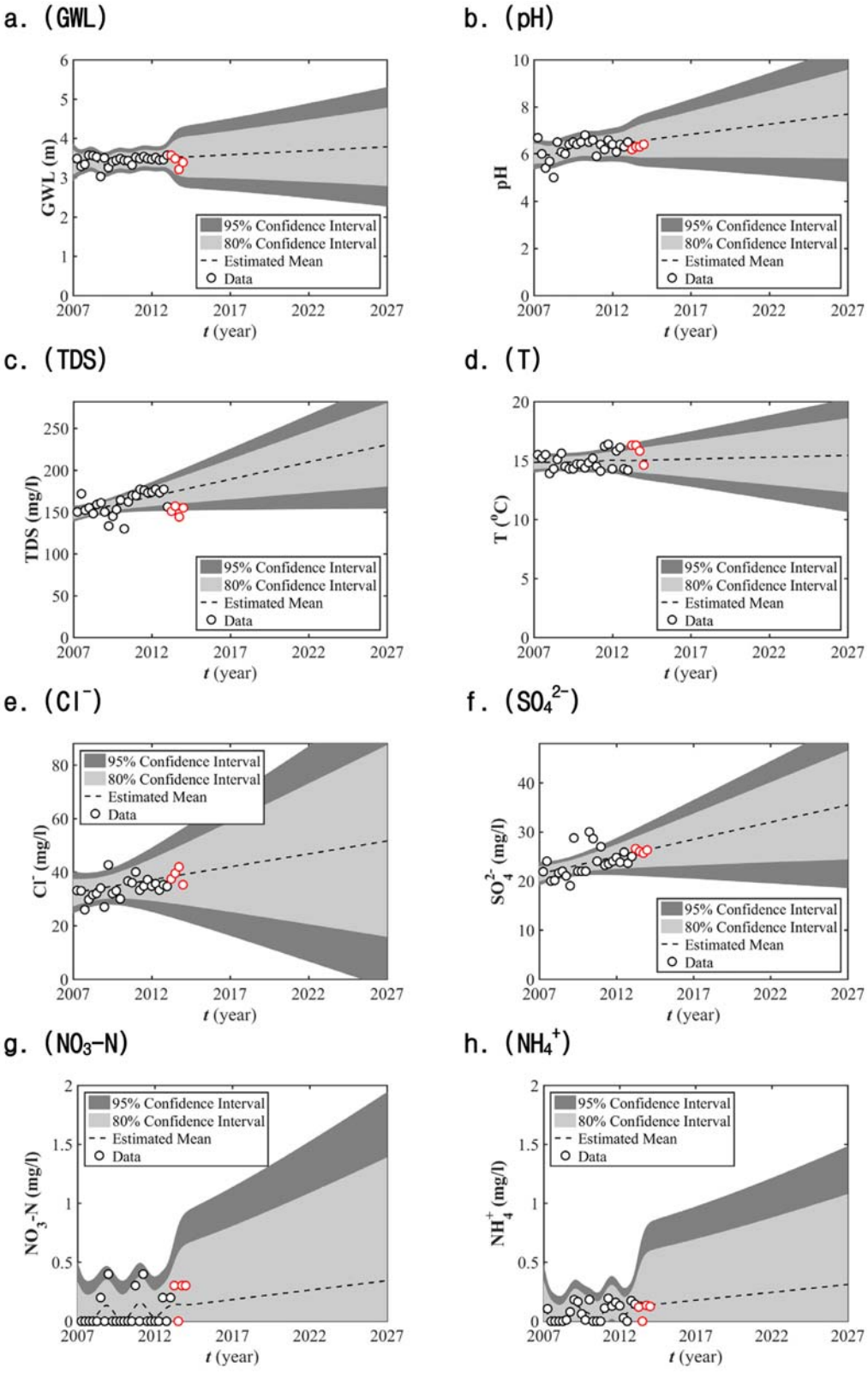

Fig. 5. Trend estimations for groundwater level (GWL), pH, Total dissolved solid (TDS), temperature (T), Cl^-, SO_4^{2-}, NO_3-N, and NH_4^+ data in Kangreung-Janghyeon monitoring location. Black circles, red circles, black dashed-lines, dark- and light-gray parcels indicate training dataset, test dataset, the estimated mean trends and the 95% and 80% confidence intervals, respectively.

지표로부터 10 m이며 관측 이래 7년 동안의 평균 지하수 심도는 3.4 m에 형성되어 있다. Fig. 5는 앞서와 동일한 8개 관측항목에 대한 GPR을 수행하고 2027년까지의 예측을 실시한 결과이다.

Fig. 5a에서 살펴 볼 수 있는 바와 같이 지하수 심도의 경우 매우 미약한 상승 추세가 관찰되며 예측된 평균선은 2007년을 기점으로 하여 $GWL = 3.38 + 0.020t$와 같다. 그러나 2027년 예측불확실성 지표(95% 신뢰구간 크기의 절반)인 2.16 m와 비교해 보았을 때 20년간의 지하수위 변동량은 18.6%이며 추세의 신뢰성이 비교적 높지 않은 편이다. 잔차들의 정규성을 테스트하기 위한 방법으로 본 연구에서는 잔차들의 분포와 이상적인 정규분포 간의 왜도 및 첨도 부합 정도를 통해 자료의 정규성을 가늠하는 Jarque-Bera 테스트(Jarque and Bera, 1987)(이하 JB 정규성 테스트)를 이용하였다. JB 정규성 테스트에 의하면 예측된 추세와 관측자료에 의한 잔차가 비정규성을 따른다고 볼 수 있다.

Fig. 5b는 pH의 20년간에 걸친 예측 추세를 보여주며 pH의 경우 앞선 지하수 심도에 비해 상대적으로 급격한 상승 추세가 관찰된다. 예측된 평균선은 $pH = 5.98 + 0.086t$와 같다. 2027년 예측불확실성 지표인 4.10과 비교해 보았을 때 20년간의 pH 변화는 41.8%이며 추세의 신뢰성이 비교적 높아 상승 추세가 나타나고 있다고 판단할 수 있다. JB 정규성 테스트에 의하면 예측된 추세와 관측자료에 의한 잔차가 정규성을 따른다.

Fig. 5c는 TDS의 20년간 예측 추세를 보여주며 TDS의 경우 비교적 큰 상승 추세가 관찰된다. 예측된 평균선은 $TDS = 147.81 + 4.127t$와 같다. 2027년 예측불확실성 지표인 108.88 mg/l와 비교해 보았을 때 20년간의 TDS 변화량은 82.53 mg/l로 75.8%이며 추세의 신뢰성이 매우 높아 뚜렷한 상승 추세가 일어나고 있다고 판단할 수 있다. 잔차들을 이용한 JB 정규성 테스트에 의하면 잔차의 분포가 비정규성을 따른다.

Fig. 5d는 GPR에 의한 지하수 수온의 20년간 예측 추세를 보여준다. 온도의 경우 앞선 지하수 심도와 마찬가지로 매우 미약한 상승 추세가 관찰되며 예측된 평균선은 $T = 14.81 + 0.031t$와 같다. 2027년 예측불확실성 지표인 6.85°C와 비교해 보았을 때 20년간의 온도 변화는 0.63°C로 10% 이하의 미미한 수준이므로 이러한 추세에 큰 의미를 부여하기는 어려우며 온도의 변화는 매우 작다고 할 수 있다. JB 정규성 테스트에 의하면 잔차의 분포가 정규성을 따른다.

Fig. 5e는 지하수 내 염소이온의 20년간 예측 추세이다. 염소이온의 경우 비교적 큰 상승 추세로 분석되며, 예측된 평균선은 $Cl^- = 32.29 + 0.966t$와 같다. 2027년 예측불확실성 지표인 78.09 mg/l와 비교해 보았을 때 20년간의 염소이온 농도변화는 19.32 mg/l로 24.7%이며 추세의 신뢰성이 비교적 높아 상승 추세가 일어나고 있다고 판단할 수 있다. JB 정규성 테스트에 의하면 잔차의 분포가 비정규성을 따른다.

Fig. 5f는 GPR에 의한 황산이온의 20년간 예측 추세를 보여주며 황산이온의 경우 역시 비교적 큰 상승 추세가 관찰된다. 예측된 평균선은 $SO_4^{2-} = 21.47 + 0.699t$와 같다. 2027년 예측불확실성 지표인 24.12 mg/l와 비교해 보았을 때 20년간의 지하수 내 황산이온 변화량은 13.99 mg/l로 58.0%이며 추세의 신뢰성이 매우 높아 뚜렷한 상승 추세가 일어나고 있다고 판단할 수 있다. JB 정규성 테스트 결과 잔차의 분포가 비정규성을 따른다.

Fig. 5g는 질산성질소의 20년간 예측 추세를 보여준다. 질산성질소의 경우 비정규성이 매우 커 GPR을 통한 분석이 부적절한 것으로 판단된다.

Fig. 5h는 암모니아성 질소의 20년간 예측 추세를 보여준다. 암모니아성 질소의 경우 완만한 상승 추세가 관찰되며, 예측된 평균선은 $NH_4^+ = 0.04 + 0.013t$와 같다. 2027년 예측불확실성 지표인 1.67 mg/l와 비교해 보았을 때 20년간의 지하수 내 암모니아성 질소 변화량은 0.27 mg/l로 16.0%이며 추세의 신뢰성이 비교적 높지 않다. JB 정규성 테스트에 의하면 잔차의 분포가 정규성을 따른다.

4.2. 평창유천 관측소

강릉장현 관측소는 강원도 평창군 대관령면 유천리 747-10 유천보건진료소 내에 위치하고 있으며 측정심도의 매질은 화강암으로 이루어져 있다. 관측소의 최상부 관측심도는 지표로부터 10 m이며 관측 이래 7년 동안의 평균 지하수 심도는 4.3 m에 형성되어 있다. Fig. 6은 앞서와 동일한 8개 관측항목에 대한 GPR을 수행하고 2027년까지의 예측을 실시한 결과이다. 평창유천 관측소의 분석항목 중 TDS, Cl^-, 및 NO_3-N의 경우 타 관측소와는 달리 증가 후 감소하는 경향을 매우 뚜렷하게 보여주고 있으며 이에 대한 추가적인 분석을 실시하였다.

Fig. 6a에서 살펴 볼 수 있는 바와 같이 지하수 심도의 경우 매우 미약한 상승 추세가 관찰되며 예측된 평균선은 2007년을 기점으로 하여 $GWL = 4.22 + 0.011t$와 같다. 그러나 2027년 예측불확실성 지표인 3.08 m와 비교해 보았을 때 20년간의 지하수위 변동량은 10% 이하의 미미

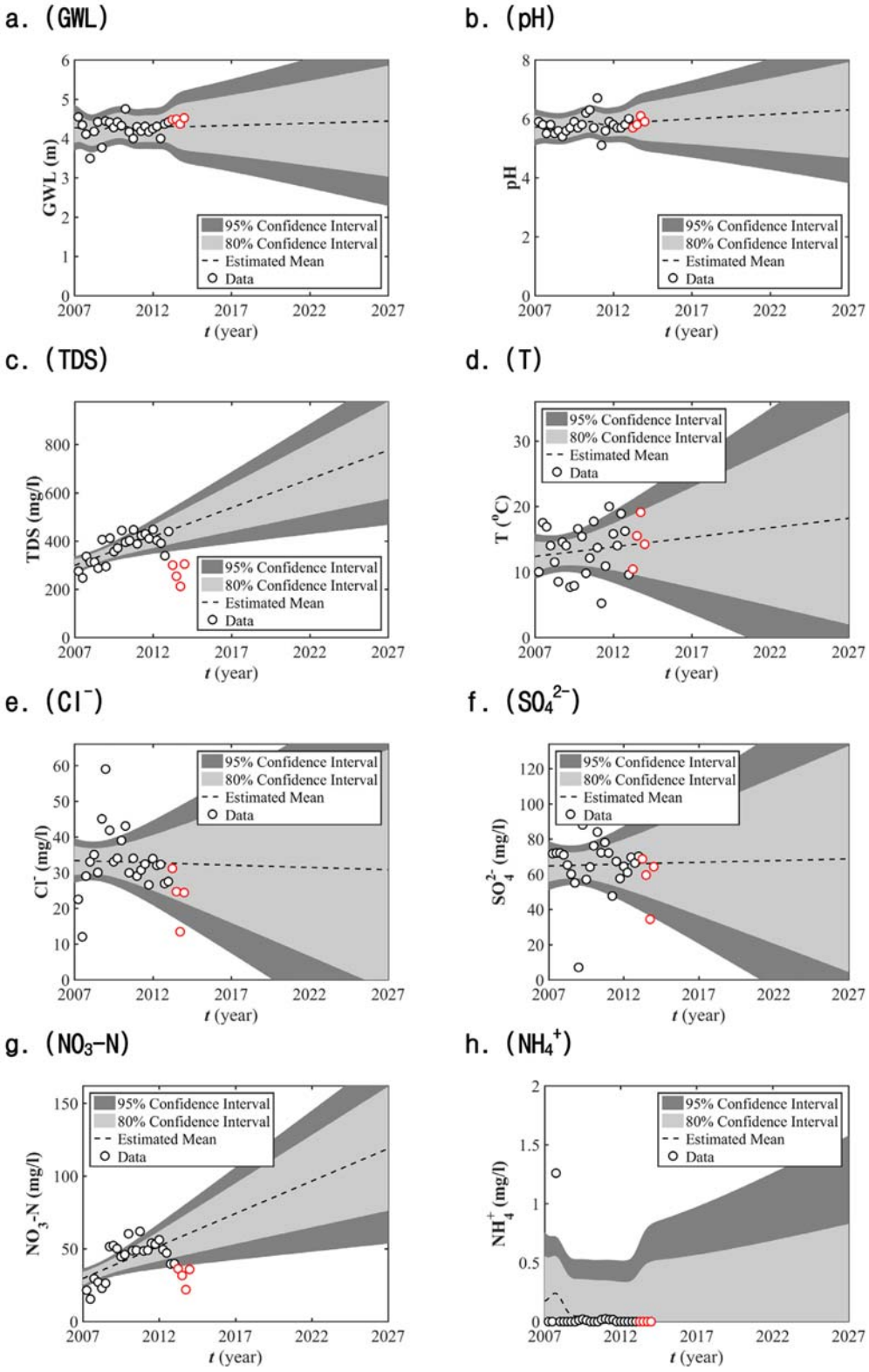

Fig. 6. Trend estimations for groundwater level (GWL), pH, Total dissolved solid (TDS), temperature (T), CL^-, SO_4^{2-}, NO_3-N, and NH_4^+ data in Pyeongchang-Yuchyeon monitoring location based on the 1st order linear regression. Black circles, red circles, black dashed-lines, dark- and light-gray parcels indicate training dataset, test dataset, the estimated mean trends and the 95% and 80% confidence intervals, respectively.

한 수준이므로 이러한 추세에 큰 의미를 부여하기는 어렵다. JB 정규성 테스트에 의하면 예측된 추세와 관측자료에 의한 잔차가 비정규성을 따른다고 볼 수 있다.

Fig. 6b는 pH의 20년간에 걸친 예측 추세를 보여주며 pH의 경우 역시 매우 미약한 상승 추세가 관찰된다. 예측된 평균선은 $pH = 5.69 + 0.031t$와 같다. 2027년 예측불확실성 지표인 3.53과 비교해 보았을 때 20년간의 pH 변화는 17.4%이며 추세의 신뢰성이 비교적 높지 않다. JB 정규성 테스트에 의하면 잔차가 비정규성을 따른다.

Fig. 6c는 GPR에 의한 TDS의 20년간 예측 추세를 보여주며 TDS의 경우 앞선 두 관측항목과 달리 상대적으로 큰 상승 추세가 관찰된다. 예측된 평균선은 $TDS = 298.97 + 23.85t$와 같다. 그러나 검증자료인 붉은색 원과 평균 예측선을 비교하여 보았을 때 예측된 평균선 및 95% 신뢰구간은 2014년 TDS 값을 예측하는데 적절하지 못한 것으로 판단된다. 이에 따라 2차식을 예측평균으로 이용한 GPR을 실시하였으며 그 결과는 Fig. 7a와 같다. 2차식에 기초한 GPR 분석결과 예측된 평균선은 $TDS = 238.33 + 79.83t - 8.96t^2$이며, 실제 2012년을 기점으로 증가하던 TDS가 감소 경향으로 변화하는 추세를 잘 반영한다. 따라서 평창유천 TDS와 같이 상승 및 하강이 동시에 관찰되는 수질자료의 추세분석을 위하해서는 2차식을 예측평균선으로 이용하는 것이 보다 타당함을 의미한다. 2027년 예측불확실성 지표인 2352.0 mg/l와 비교해 보았을 때 20년간의 TDS 변화량은 1596.6 mg/l로 67.9%로 분석되었으나 추세가 선형의 변화를 보여주지 않으므로 상승 또는 하강 추세 분석은 무의미하다. 선형 및 비선형 예측결과에 의한 잔차들의 분포는 JB 정규성 테스트에 의해 모두 정규성을 따르는 것으로 분석되었다.

Fig. 6d는 지하수 수온의 20년간 예측 추세를 보여준다. 온도의 경우 비교적 큰 상승 추세가 관찰된다. 예측된 평균선은 $T = 12.36 + 0.291t$와 같다. 2027년 예측불확실성 지표인 35.4ºC와 비교해 보았을 때 20년간의 온도 변화는 5.82ºC로 16.5%이며 추세의 신뢰성이 비교적 높지 않다. JB 정규성 테스트에 의하면 잔차의 분포가 정규성을 따른다.

Fig. 6e는 지하수 내 염소이온의 20년간 예측 추세이다. 염소이온의 경우 매우 미약한 하강 추세로 분석되며, 예측된 평균선은 $Cl^- = 33.38 - 0.129t$와 같다. 그러나 검증자료와 예측결과를 보았을 때 예측된 평균선 및 95% 신뢰구간은 염소이온 값을 예측하는데 적절하지 못하다. 이에 따라 앞선 TDS와 동일하게 2차식을 예측평균으로 이용한 GPR을 실시하였으며 그 결과는 Fig. 7b와 같다. 예측된 평균선은 $Cl^- = 21.22 + 11.09t - 1.8t^2$이며, 실제 2010년을 기점으로 증가하던 염소이온이 감소 경향으로 변화하는 추세를 잘 반영한다. 2027년 예측불확실성 지표인 445.96 mg/l와 비교해 보았을 때 20년간의 염소이온 농도 변화는 221.85 mg/l로 49.7%로 분석되었으나 2차식을 적용함으로 인하여 상승 또는 하강 추세를 단정할 수는 없다. 예측된 추세와 관측자료 간의 잔차들을 이용한 JB 정규성 테스트에 의하면 잔차의 분포가 1 및 2차식 모두에서 비정규성을 따른다.

Fig. 6f는 GPR에 의한 황산이온의 20년간 예측 추세를 보여주며 황산이온의 경우 매우 미약한 상승 추세가 관찰된다. 예측된 평균선은 $SO_4^{2-} = 64.74 + 0.196t$와 같다. 2027년 예측불확실성 지표인 140.31 mg/l와 비교해 보았을 때 20년간의 지하수 내 황산이온 변화량은 3.93 mg/l로 10% 이하의 미미한 수준이므로 이러한 추세에 큰 의미를 부여하기는 어렵다. JB 정규성 테스트에 의하면 잔차의 분포가 비정규성을 따른다.

Fig. 6g는 질산성질소의 20년간 예측 추세를 보여주며, 질산성질소의 경우 비교적 큰 상승 추세가 관찰된다. 예측된 평균선은 $NO_3\text{-}N = 29.51 + 4.461t$와 같다. 그러나 검증자료와 예측결과를 보았을 때 예측된 평균선 및 95% 신뢰구간은 질산성질소 값을 예측하는데 적절하지 못하다. 따라서 2차식을 예측평균으로 이용한 GPR을 실시하였으며 그 결과는 Fig. 7c와 같다. 예측된 평균선은 $NO_3\text{-}N = 9.13 + 23.27t - 3.01t^2$이며, 실제 2011년을 기점으로 증가하던 질산성질소가 감소 경향으로 변화하는 추세를 잘 반영한다. 2027년 예측불확실성 지표인 412.92 mg/l와 비교해 보았을 때 20년간의 질산성질소 변화량은 465.48 mg/l로 112.7%로 분석되었으나 추세가 선형의 변화를 보여주지 않으므로 상승 또는 하강 추세를 단정할 수는 없다. 평창유천의 경우 2008년 관측을 시작한 이후 지속적으로 MCL을 초과하는 지하수 내 질산성질소 농도가 관측되고 있다. 그러나 2차식의 예측 평균선을 고려하여 볼 때 2007년 이후 8년에 해당하는 2016년경에는 예측 평균선이 질산성질소의 MCL인 10 mg/l를 만나게 되어 2016년 이후부터는 MCL 이하의 지하수 내 질산성질소 농도가 관측될 가능성이 있으며, 2017년 말경 95% 신뢰구간과 MCL이 교차하여 이 시기 이후 질산성질소에 의한 지하수 오염이 사라질 가능성이 높은 것으로 분석된다. JB 정규성 테스트에 의하면 잔차의 분포가 정규성을 따른다.

Fig. 6h는 암모니아성 질소의 20년간 예측 추세를 보여며 자료의 비정규성이 매우 커 GPR을 통한 분석이 부적절한 것으로 판단된다.

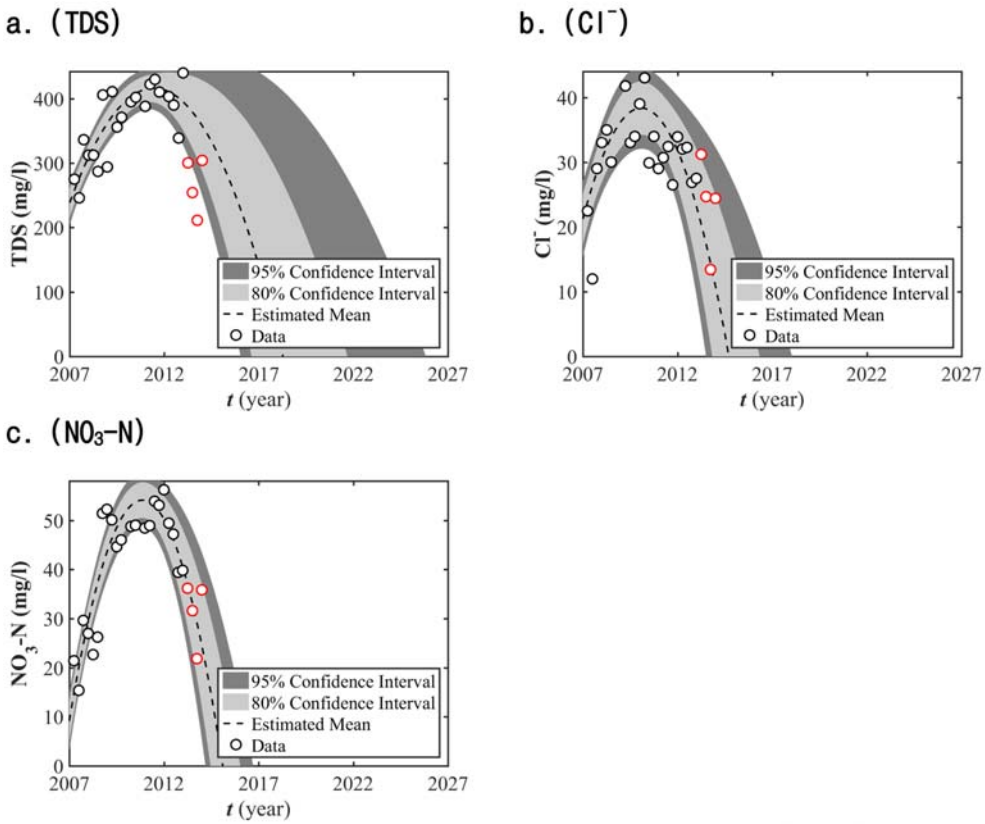

Fig. 7. Trend estimations for Total dissolved solid (TDS), Cl^-, and NO_3-N data in Pyeongchang-Yuchyeon monitoring location based on the 2nd order linear regression. Black circles, red circles, black dashed-lines, dark- and light-gray parcels indicate training dataset, test dataset, the estimated mean trends and the 95% and 80% confidence intervals, respectively.

5. 결 론

본 연구에서는 모수적 지하수질 추세분석 기법으로 가우시안 프로세스 회귀분석기법(GPR)을 제시하였다. GPR 기법은 수질 추세에 대한 예측과 동시에 예측에 따른 불확실성을 제시하여준다는 측면에서 매우 유용한 방법론이라고 할 수 있다. 또한 예측평균선의 비선형성과 주기적인 변화 등을 활용하는데 제한성이 없어 기존 비모수적 방법의 한계를 극복할 수 있을 것으로 판단된다. 그러나 GPR은 근본적으로 선형 회귀분석에 근거하기 때문에 1차식을 예측평균으로 하는 선형 추세 예측에 있어서 지하수 관측항목 중 이상치에 대한 민감성을 보이며, 따라서 Theil-Sen 추정에 비하여 불안정한 예측을 한다. 또한 예측 추세에 의한 잔차들이 정규분포에서 크게 벗어나 있는 추세를 설명할 경우 이론적 기반이 취약할 수 있다. 이러한 GPR의 장점과 단점을 설명하기 위하여 GPR의 기반이 되는 일반 선형 회귀분석과 Theil-Sen 추정을 가상의 자료를 통하여 설명하였으며, 특히 비선형성이 심각한 자료의 경우 모수적 방법의 취약성을 설명하였다.

GPR의 수질추세 예측 적용성을 살펴보기 위하여 7년 이상의 관측이 이루어진 지하수 수질전용측정망 2개소(강릉장현 및 평창유천)에 대한 적용 및 분석이 이루어졌으며 분석 수질항목은 총 8개(지하수위, pH, TDS, T, 염소이온, 황산이온, 질산성질소, 암모니아성질소)이다. 분석에 활용된 전체기간인 총 7년(28개 분기) 중 6년(24개 분기)의 자료를 이용하여 GPR을 훈련시킨 후 최종 1년(4개 분기)의 자료를 검증자료로 이용하여 예측의 건전성을 시험하였다. 전반적으로, 비선형 추세를 지니는 관측 수질항목을 제외하면, 1차식을 예측평균선으로 이용한 예측에서 대부분의 검증자료가 신뢰구간 내에 분포하는 것으로 분석되었으며, 따라서 건전한 예측이 이루어졌다고 판단할 수 있다.

평창유천의 경우에는 TDS, 염소이온, 질산성질소가 뚜렷한 증가 후 감소 경향을 보여주고 있다. TDS의 경우 2011년 3월 경 증감추세가 바뀌며, 염소이온의 경우 2010년 1월 경 그리고 질산성질소의 경우 2010년 11월 경

증감추세가 교차하는 것으로 예측되었다. 또한 질산성질소의 경우 2017년 말경 95% 신뢰구간과 MCL이 교차하는 것으로 분석되어, 2014년 현재 자료를 통해 판단할 때, 이 시기 이후 질산성질소에 의한 지하수 오염이 사라질 수 있는 분석된다.

사 사

이 논문은 2013학년도 경북대학교 전임교원 연구년 교수 연구비에 의하여 연구되었음

References

Kim, G.B., Choi, D.H., Yoon, P.S., and Kim, K.Y., 2010, Trends of groundwater quality in the areas with a high possibility of pollution, *J. Korean Geo-Environ. Soc.*, **11**(3), 5-16.

Ministry of Environment (Korea), National Institute of Environmental Research (Korea), 2007-2013, National Groundwater Quality Monitoring Network Annual Report.

Bazi, Y., Alajlan, N., and Melgani, F., 2012, Improved Estimation of Water Chlorophyll Concentration With Semisupervised Gaussian Process Regression, *IEEE Trans. Geosci. Remote Sensing*, **50**(7), 2733-2743.

Chapman, D., 1996, Water quality assessments: a guide to the use of biota, sediments, and water in environmental monitoring, *UNESCO/WHO/UNEP*, 22 p.

Grbić, R., Kurtagić, D., and Slišković, D., 2013, Stream water temperature prediction based on Gaussian process regression, *Expert Sys. Applic.*, **40**(18), 7407-7414.

Helsel, D.R. and Hirsch, R.M., 1988, Applicability of the t-Test for Detecting Trends in Water Quality Variables, *J. American Water Resour. Assoc.*, **24**(1), 201-204.

Helsel, D.R. and Hirsch, R.M., 2002, Statistical methods in water resources: US Geological Survey Techniques of Water Resources Investigations, book 4, chap. A3, U.S. Geological Survey.

Hirsch, R.M., Slack, J.R., and Smith, R.A., 1982, Techniques of trend analysis for monthly water-quality data, *Water Resour. Res.*, **18**, 107-121.

Hirsch, R.M., Alexander, R.B., and Smith, R.A., 1991, Selection of methods for the detection and estimation of trends in water quality, *Water Resour. Res.,* **27**(5), 803-813.

Jarque, Carlos M., Bera, Anil K., 1987, A test for normality of observations and regression residuals, *Int. Stat. Rev.*, **55**(2), 163-172.

Murphy, K.P., 2012, Machine Learning: a Probabilistic Perspective, The MIT Press, Cambridge, 1067 p.

Sun, A.Y., Wang, D., and Xu, X., 2014, Monthly streamflow forecasting using Gaussian Process Regression, *J. Hydro.*, **511**, 72-81.