과업 수행계획서

과 제 명 : 딥러닝 기법을 활용한 4대강 수계 주요 유입지류 수질변화 연구

연구기간 : 2021년 10월 5일~2022년 6월 2일
(계약일로부터 8개월)

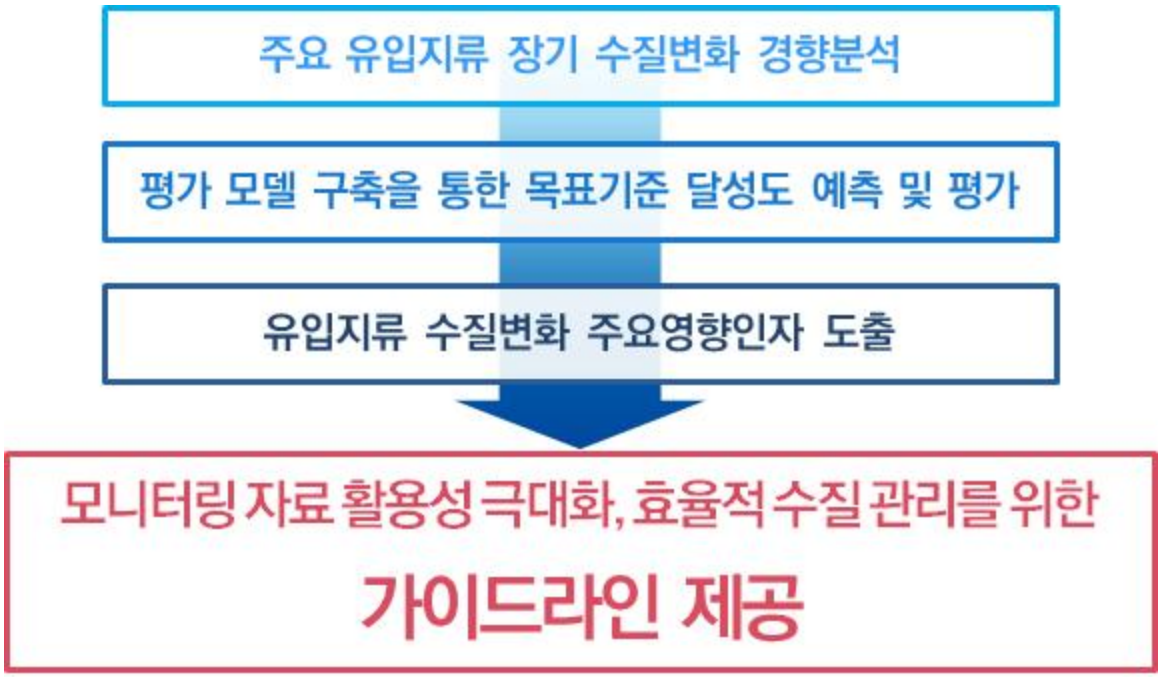
주 관 기 관 : 국립환경과학원
용역수행기관 : 서울시립대학교
연구책임자 : 서울시립대학교 차윤경

1. 과업수행 방향 및 방법

1-1. 과업의 목적

- 본 과업에서는 딥러닝 및 조건부확률모델과 기작기반 모델의 융합적 활용을 통해 본류의 수질에 큰 영향을 미치는 유입지류에 대해 외부영향으로 인한 수질 변화 분석을 목적으로 하며, 분석 결과를 바탕으로 중점 관리가 필요한 주요 유입지류 지점 및 지점별 수질항목을 선정하여 지속적인 노력을 통해 축적된 물환경 모니터링 자료의 활용성을 극대화함과 동시에 효율적인 유입지류 수질관리를 위한 가이드라인을 제공하고자 한다 <그림 1-1>.

- 이를 위한 세부 목표는 아래와 같이 3 가지로 구성했다.
  - ✓ 주요 유입지류 장기 수질변화 경향 분석
  - ✓ 평가 모델(딥러닝 모델 활용) 구축을 통한 목표기준 달성도 예측 및 평가
  - ✓ 유입지류 수질변화 주요영향인자 도출



<그림 1-1> 과업의 목적 및 세부목표

1-2. 과업 수행 방법

○ 주요 유입지류 장기 수질변화 경향 분석

- 국내 5대강 수계의 본류 및 지류의 물환경 모니터링 지점에 대한 물리, 화학적 수질항목, 기상, 유량, 유속 등 물 흐름 등 가용 모니터링 자료를 수집한다.
- 수집된 물환경 모니터링 자료를 기반으로 모니터링 지점의 수질 특성 등을 분석, 분석결과를 바탕으로 연구 대상지역을 선정한다.
- 수집된 물환경 모니터링 자료에 대한 기초자료 분석 등을 통해 연구 대상지역에서 강수량 등 기상 변화와 상·하류 등 공간적 차이 등에 따른 수질항목(BOD, T-P, TOC 등)의 현황과 중권역 대표지점에 대해 목표 기준 달성여부 현황을 조사하고 달성여부에 대한 추세 등을 분석한다.

○ 평가 모델 구축을 통한 목표기준 달성도 예측 및 평가

- 연구 대상지역의 수질 특성에 따라 최적 입력변수를 선정하고 가용 자료 현황 및 가용 모델링 기법 등을 검토, 최적 모델을 선정한다.
- SWAT 등 기작모델을 활용하여 딥러닝 및 조건부확률 모델링을 위한 추가 입력자료를 생산한다.
- 모니터링 자료와 추가 입력자료를 활용하여 딥러닝 모델을 구축하고 BOD, T-P, TOC 등에 대한 목표기준 달성도 및 외부영향에 따른 수질변화 등을 예측한다.
- 모니터링 자료와 추가 입력자료를 활용하여 조건부확률 모델을 구축하고 BOD, T-P, TOC 등에 대한 목표기준 달성도 및 외부영향에 따른 수질변화 등을 예측한다.

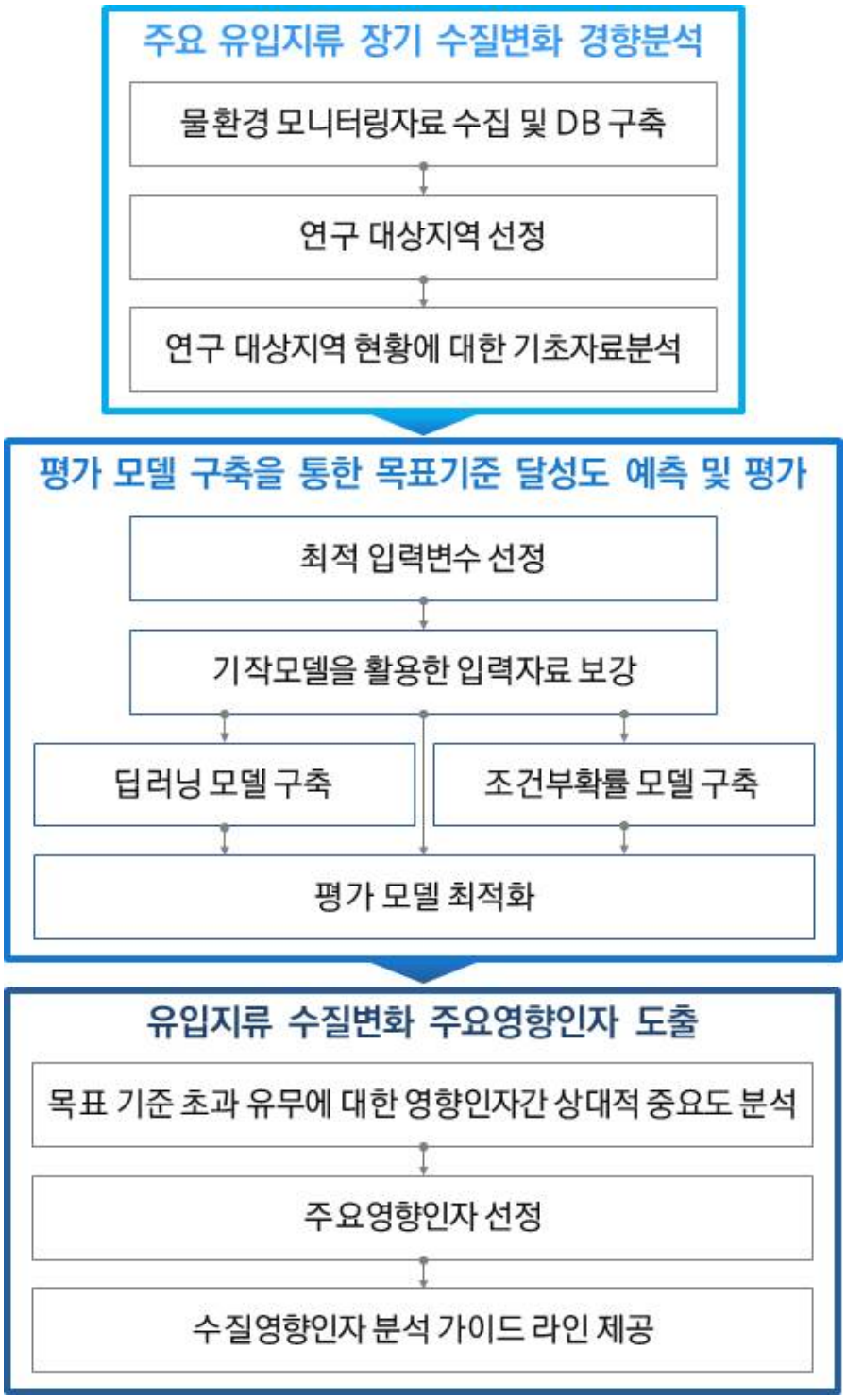
○ 유입지류 수질변화 주요영향인자 도출

- 구축된 딥러닝 모델에 대한 모델 해석 기법 등을 적용하여 연구 대상지역의 주요 수질영향인자를 도출한다.
- 구축된 조건부확률 모델을 활용하여 연구 대상지역에의 주요 수질영향인자를 도출한다.
- 주요 수질영향인자 도출을 위한 사용자 매뉴얼 등을 작성한다.

2. 과업내용 및 세부 수행계획

2-1. 과업내용

- 본 연구는 유입지류 수질변화에 대한 주요영향인자 도출을 목적으로 하며, 연구 흐름도는 아래와 같다 <그림 2-1>.

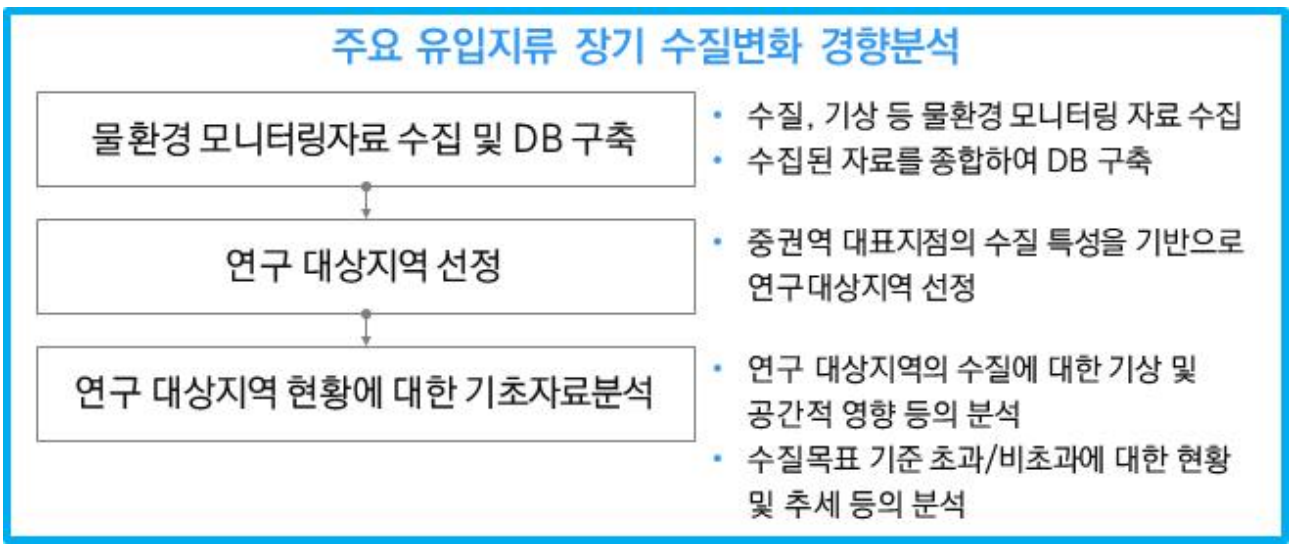


<그림 2-1> 연구 흐름도(안)

2-2. 세부 수행계획

2-2-1 주요 유입지류 장기 수질변화 경향 분석

- 딥러닝 및 조건부확률 모델 구축을 위한 데이터 수집 및 가공, 기초자료분석 단계로 물환경 모니터링자료 수집 및 DB 구축, 연구 대상지역 선정, 연구 대상지역 현황에 대한 기초자료 분석으로 구성한다 <그림 2-2>.
  - 물환경 모니터링자료 수집 및 DB 구축 단계에선 가용 모니터링 자료 현황을 파악한 후 모델 구동을 위해 요구되는 자료 및 분석을 위해 필요하다고 판단되는 항목을 수집하여 모델 구축을 위한 DB를 구축한다.
  - 예를 들어, SWAT 모델 구동을 위해서는 기상자료와 공간자료가 요구되며, 기상입력자료로는 일강수량(mm), 최고 및 최저기온(℃), 일평균 풍속(m/s), 상대습도(%), 일사량(MJ/m^2)이 사용되며, 공간입력 자료로는 토지피복도, 토양도, 수치고도모델(Digital Elevation Model, DEM)을 필요로 하므로 해당 모니터링 자료를 수집한다.
  - 연구 대상지역 선정 단계에선 수집된 DB를 기반으로 유입지류의 중권역 대표지점들의 수질 특성 등을 기반으로 연구 대상지역을 선정한다.
  - 연구 대상지역 현황에 대한 기초자료 분석 단계에선 선정된 연구 대상지역에 대해 강수량 등 기상 변화 및 상·하류에 따른 수질 차이 등에 대한 기초자료분석을 수행하고, 중권역 대표지점별 수질목표에 따라 달성/미달성에 대한 현황 및 추세 등을 분석한다.

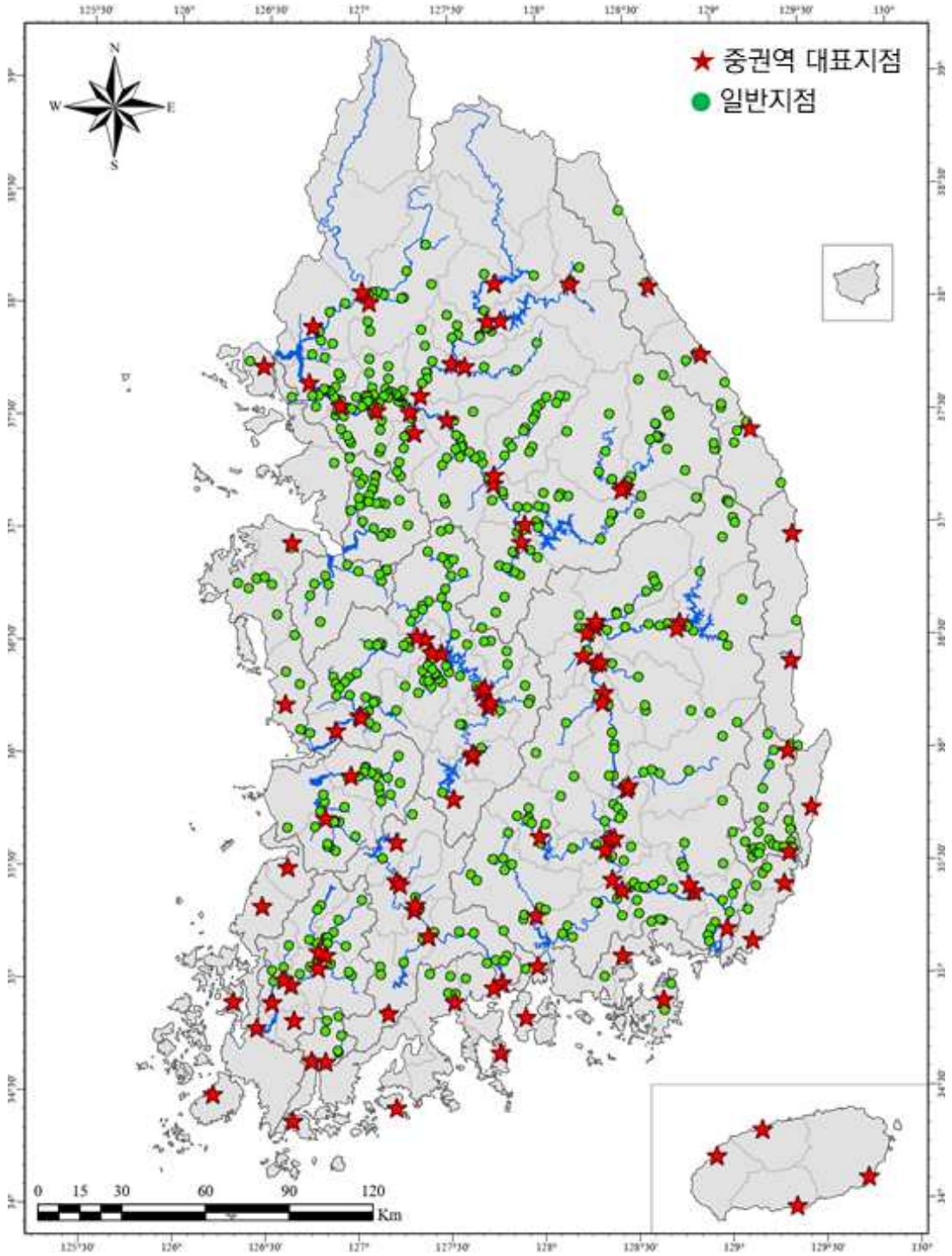


<그림 2-2> 주요 유입지류의 장기 수질변화 경향분석 흐름도(안)

2-2-1-1. 물환경 모니터링 자료 수집 및 DB 구축

✓ **하천 수질 모니터링 현황**

- 전국 5대강 수계(한강, 금강, 낙동강, 영산·섬진강) 하천 수질측정망 지점에 대한 모니터링 자료를 검토 및 수집한다 <그림 2-3>.

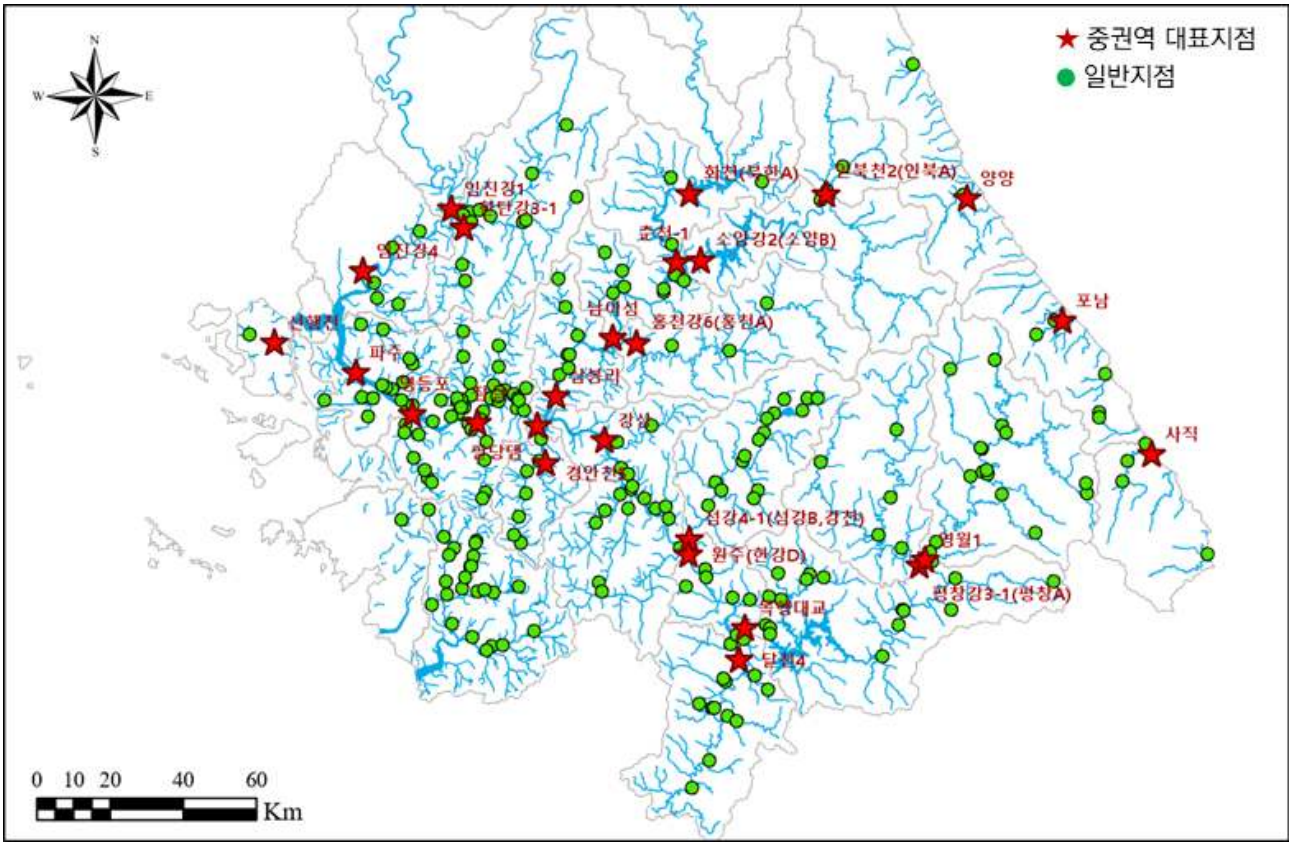


<그림 2-3> 전국 5대강 수계 수질측정망도(2021년도 물환경측정망 운영계획)

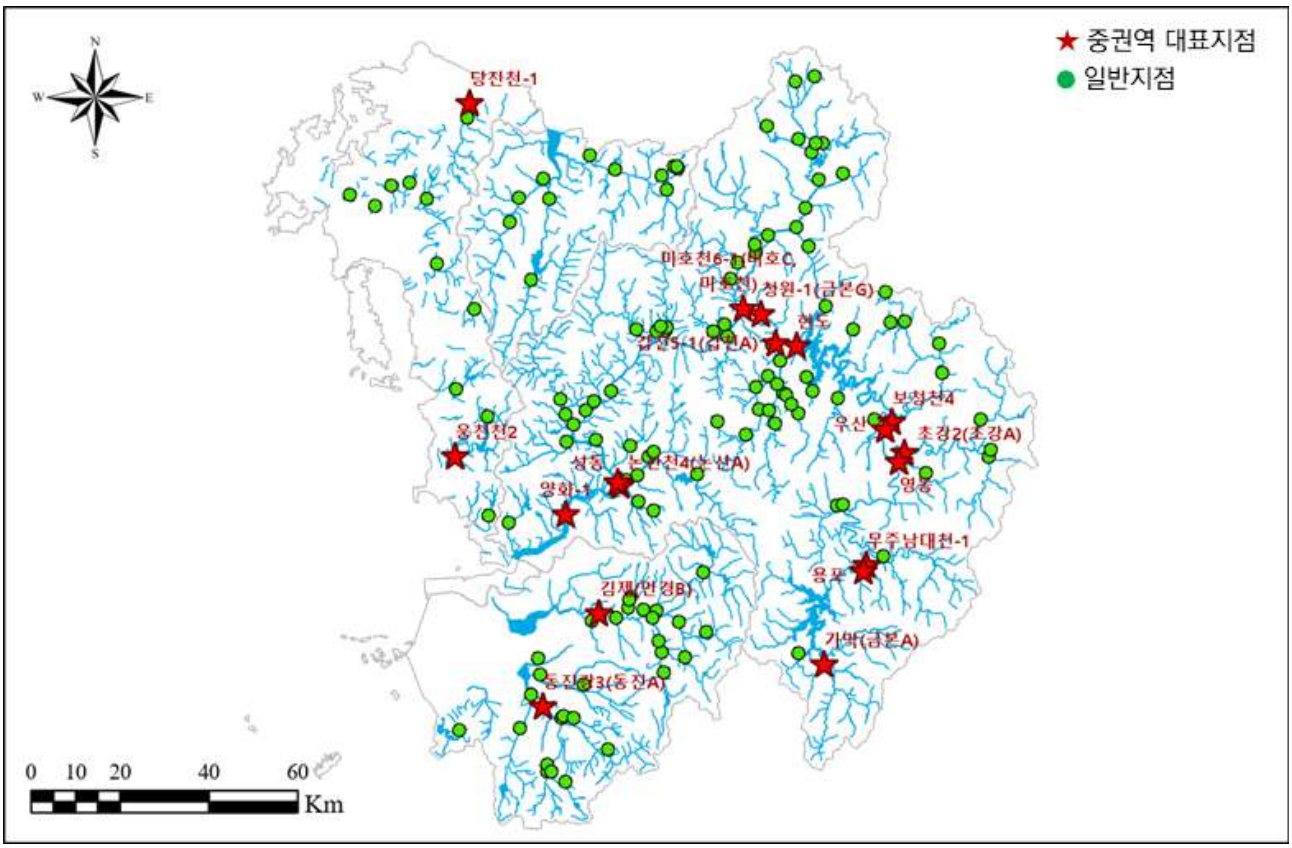
- 2021년 물환경측정망 설치운영계획을 바탕으로 파악한 하천 수질측정망 개소수는 총 688개소이며, 이중 중권역 대표지점은 총 110개소이다 <표 2-1>, <그림 2-4>, <그림 2-5>, <그림 2-6>, <그림 2-7>.
  - 총량측정망 및 호소 수질측정망은 계수에서 제외했다.
  - 총량측정망 중 수질병행 측정지점은 포함하여 계수했다.
- 하천 수질측정망 중 일반지점은 수온(℃), pH, DO(㎎/L), BOD(㎎/L), COD(㎎/L), TOC(㎎/L), SS(㎎/L), 총질소(㎎/L)를 포함한 19개 수질항목에 대해 년 12회 (매월) 측정하고 있으며, Cd(㎎/L), CN(㎎/L)를 포함한 8개 항목에 대해서는 년 4회 (분기별) 측정하고 있다.
- 하천 수질측정망 중 일반지점은 한강 대권역 234개소, 금강 대권역 129개소, 낙동강 대권역 193개소, 영산・섬진강 대권역 87개소이다.
- 중권역 대표지점 중 본류 지점은 186개소이며, 지류지점은 438개소, 서해, 동해, 남해를 포함한 기타지점은 64개소이다.
- 하천 수질측정망 중 중권역 대표지점은 일반지점과 동일하게 수온(℃), pH, DO(㎎/L), BOD(㎎/L), COD(㎎/L), TOC(㎎/L), SS(㎎/L), 총질소(㎎/L)를 포함한 19개 수질항목에 대해 년 12회 (매월) 측정하고 있으며, Cd(㎎/L), CN(㎎/L)를 포함한 8개 항목에 대해서는 년 4회 (분기별) 측정하고 있으나, 추가로 사염화탄소(㎎/L), 1,2-디클로로에탄(㎎/L), 디클로로메탄(㎎/L)을 포함한 10개 항목에 대해서는 년 2회 (3월, 9월), PCB와 유기인은 7월에 1회, 디에틸헥실프탈레이트는 10월에 1회 측정한다.
- 하천 수질측정망 중 중권역 대표지점은 한강 대권역 27개소, 금강 대권역 18개소, 낙동강 대권역 32개소, 영산・섬진강 대권역 33개소이다.
- 중권역 대표지점 중 본류 지점은 50개소이며, 지류지점은 37개소, 서해, 동해, 남해를 포함한 기타지점은 23개소이다.

<표 2-1> 하천 수질측정망 지점 개소수 및 측정항목(2021년도 물환경측정망 운영계획)

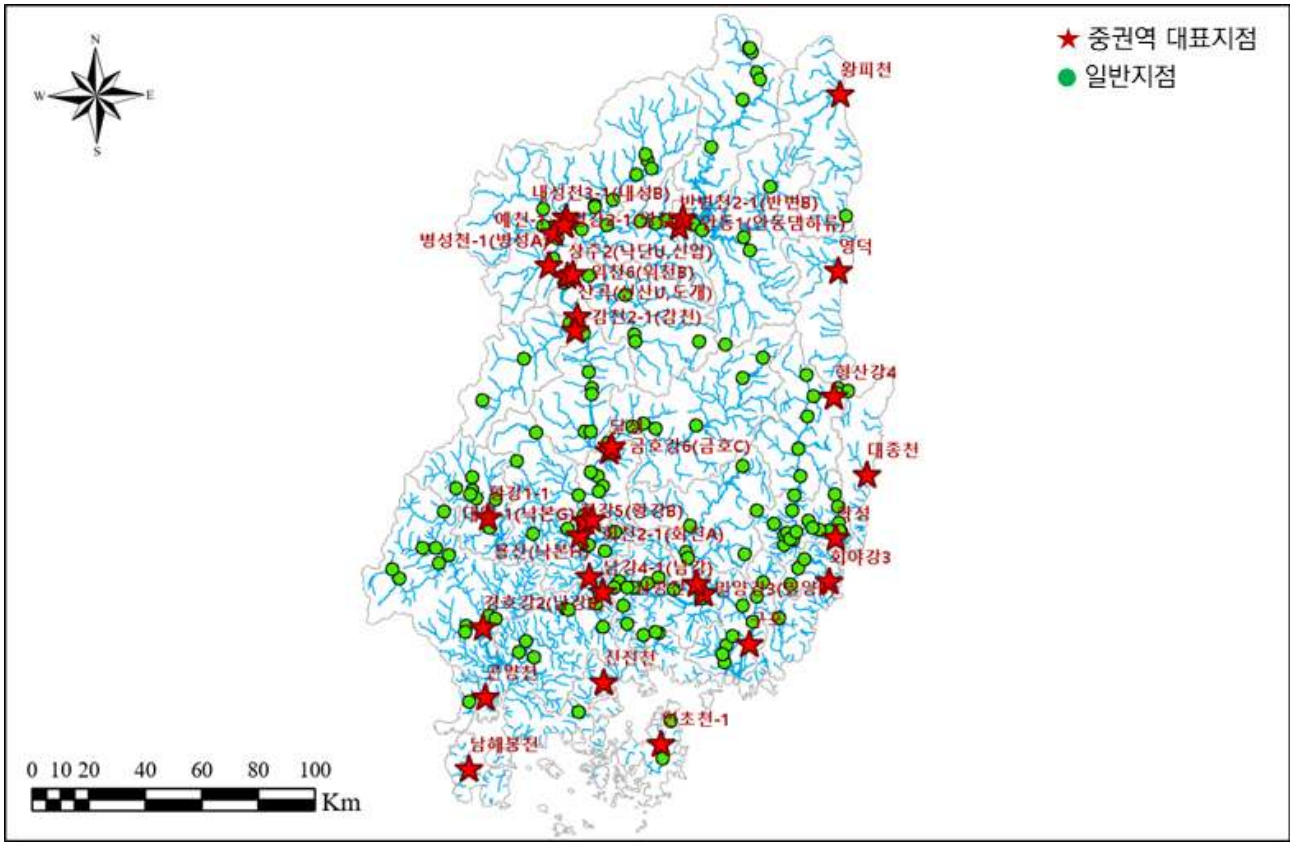
대권역	수계	개소수 (중권역 대표)	측정 항목 및 주기	
한강	남한강	101 (8)	일반지점	[12회/년(매월)] 수온(℃), pH, DO(㎎/L), BOD(㎎/L), COD(㎎/L), TOC(㎎/L), SS(㎎/L), 총질소(㎎/L), DTN(㎎/L), NH_3-N(㎎/L), NO_3-N(㎎/L), 총인(㎎/L), DTP(㎎/L), PO_4-P(㎎/L), 페놀류(㎎/L), 분원성대장균군수(분원성대장균군수/100mL), 총대장균군수(총대장균군수/100mL), 전기전도도(㎲/cm), 클로로필-a(㎎/㎥) [4회/년(3, 6, 9, 12월)] Cd(㎎/L), CN(㎎/L), Pb(㎎/L), Cr^{6+}(㎎/L), As(㎎/L), Hg(㎎/L), Sb(㎎/L), ABS(㎎/L)
	북한강	33 (7)		
	한강	83 (8)		
	안성천	26 (-)		
	한강동해	14 (3)		
	한강서해	4 (1)		
금강	금강	91 (14)		
	금강서해	13 (2)	대표지점	[12회/년(매월)] 수온(℃), pH, DO(㎎/L), BOD(㎎/L), COD(㎎/L), TOC(㎎/L), SS(㎎/L), 총질소(㎎/L), DTN(㎎/L), NH3-N(㎎/L), NO3-N(㎎/L), 총인(㎎/L), DTP(㎎/L), PO4-P(㎎/L), 페놀류(㎎/L), 분원성대장균군수(분원성대장균군수/100mL), , 총대장균군수(총대장균군수/100mL), 전기전도도(㎲/cm), 클로로필-a(㎎/㎥) [4회/년(분기별)] Cd(㎎/L), CN(㎎/L), Pb(㎎/L), Cr6+(㎎/L), As(㎎/L), Hg(㎎/L), Sb(㎎/L), ABS(㎎/L) [2회/년(3월, 9월)] TCE(㎎/L), PCE(㎎/L), 사염화탄소(㎎/L), 1, 2-디클로로에탄(㎎/L), 디클로로메탄(㎎/L), 벤젠(㎎/L), 클로로포름(㎎/L), 1, 4-다이옥세인(㎎/L), 포름알데히드(㎎/L), 헥사클로로벤젠(㎎/L) [1회/년(7월)] PCB(㎍/kg), 유기인(㎎/L) [1회/년(10월)] 디에틸헥실프탈레이트(DEHP)(㎎/L)
	만경동진	31 (2)		
	삽교천	12 (-)		
낙동강	낙동강	143 (22)		
	낙동강남해	11 (3)		
	낙동강동해	5 (3)		
	태화강	20 (1)		
	형산강	8 (1)		
	회야수영	6 (2)		
영산·섬진강	섬진강	26 (9)		
	섬진강남해	13 (7)		
	영산강	32 (8)		
	영산강남해	1 (1)		
	영산강서해	4 (3)		
	제주도	4 (4)		
	탐진강	7 (1)		
계		688 (110)		



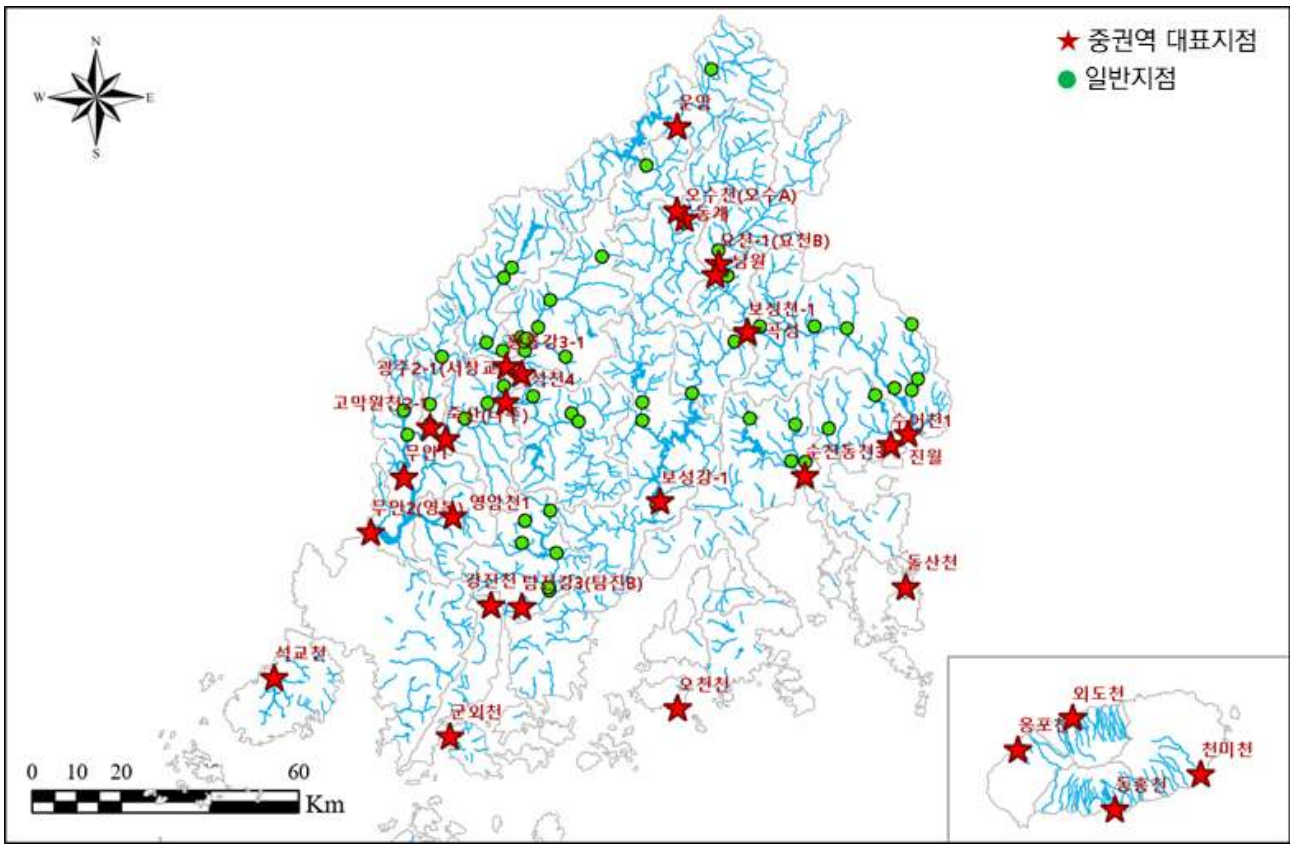
<그림 2-4> 한강 대권역 일반지점 및 중권역 대표지점



<그림 2-5> 금강 대권역 일반지점 및 중권역 대표지점



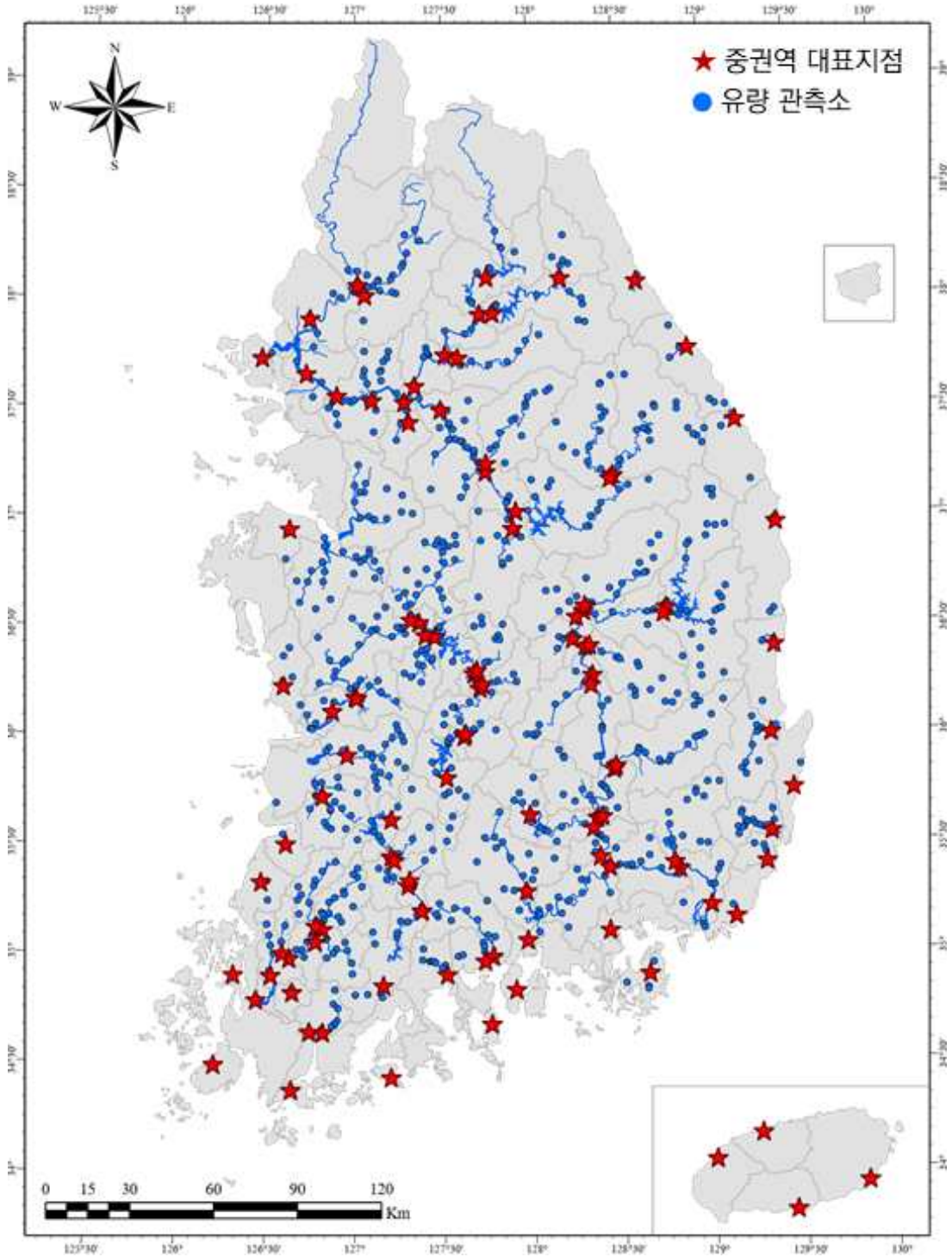
<그림 2-6> 낙동강 대권역 일반지점 및 중권역 대표지점



<그림 2-7> 영산·섬진강 대권역 일반지점 및 중권역 대표지점

✓ **유량자료 현황**

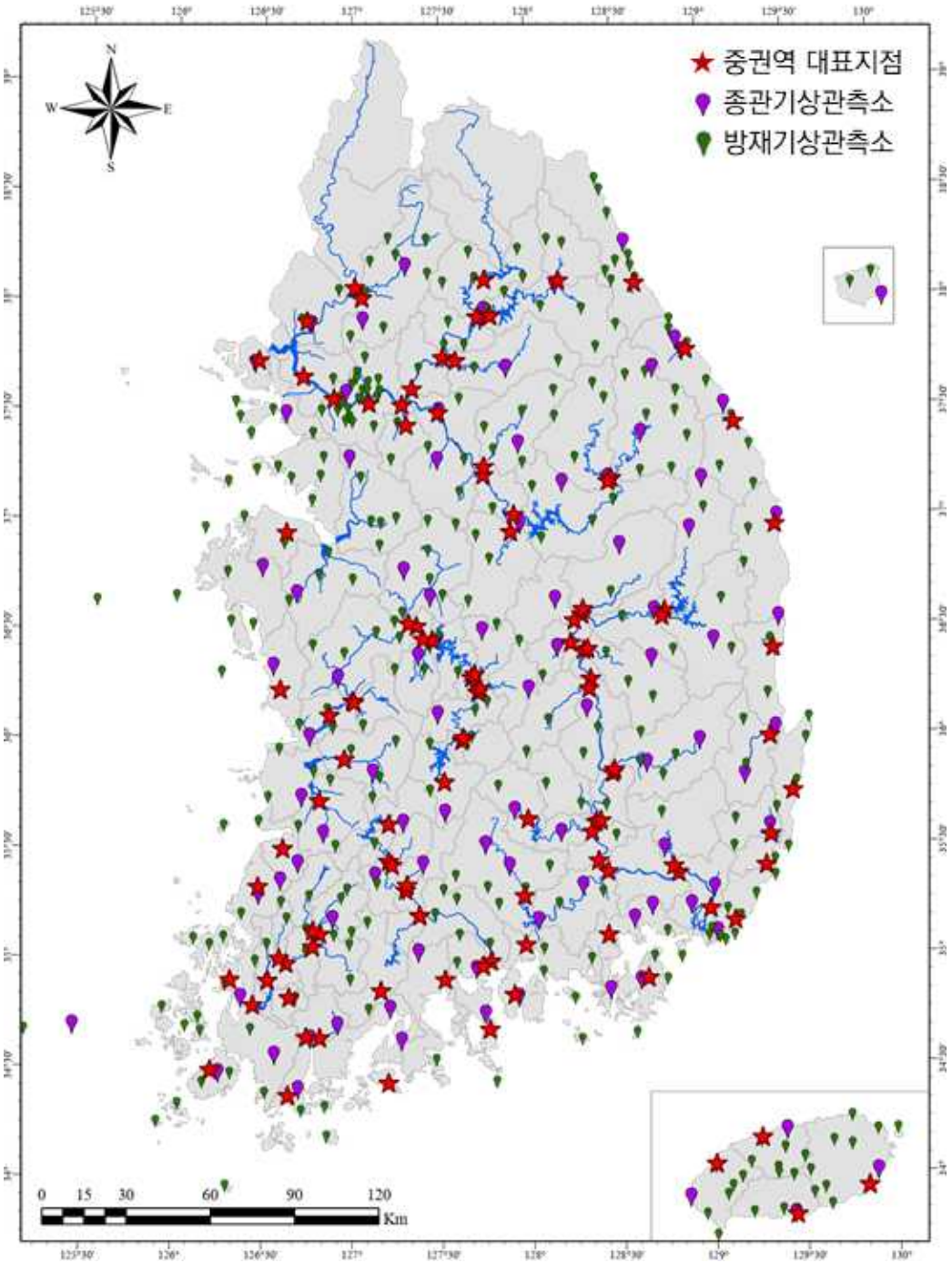
- 현재 전국에는 809개 유량관측소가 운영중에 있으며, 분포는 아래와 같다 <그림 2-8>.
- 유량관측소의 모니터링 자료는 국가수자원관리 종합정보시스템을 통해 획득할 수 있으며, 일 단위 유량자료를 제공한다.



<그림 2-8> 전국 5대강 수계 유량관측소

✓ **기상자료 현황**

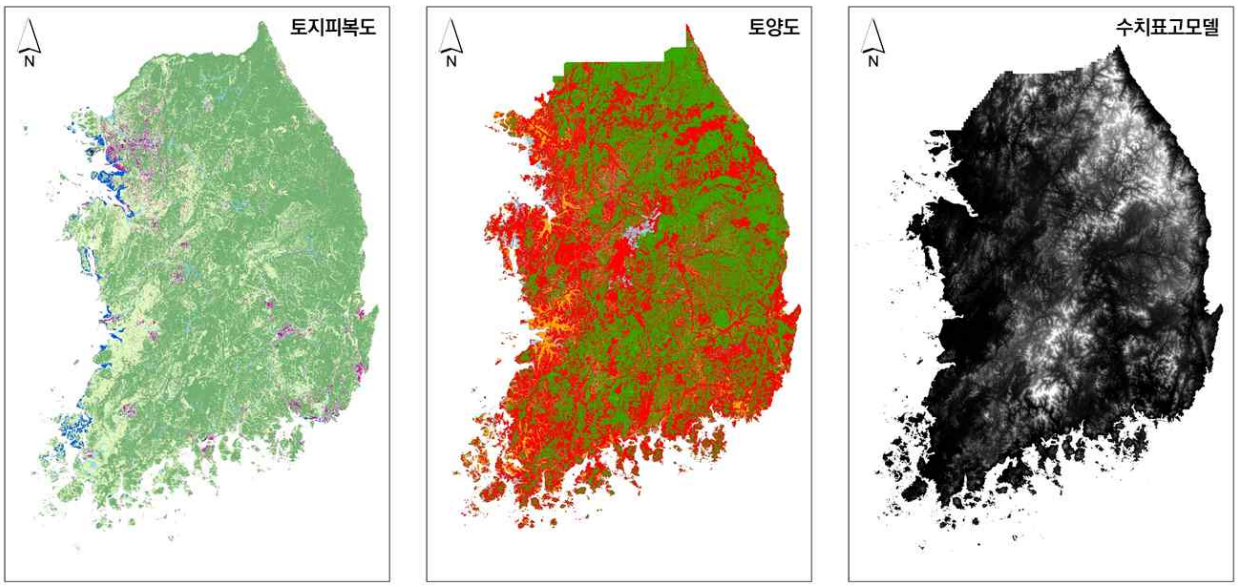
- 기상자료는 95개 종관기상관측소, 310개의 방재기상관측소의 자료를 활용할 수 있을 것으로 판단되며, 분포는 아래와 같다 <그림 2-9>.
- 기상 관측 자료는 기상청 기상자료개방포털에서 획득할 수 있으며, 일강수량(㎜), 최고 및 최저기온(℃), 일평균 풍속(m/s), 상대습도(%), 일사량(MJ/m^2) 등의 다양한 기상항목을 제공한다.



<그림 2-9> 전국 5대강 수계 유량관측소

✓ **공간자료 현황**

- 수질 변화에 대한 다각도의 분석을 위해 토지피복도, 토양도, 수치표고모델 등 다양한 공간자료를 수집한다 <그림 2-10>.
- 토지피복도는 환경부의 환경공간정보서비스를 통해 획득할 수 있으며, 대분류, 중분류, 세분류 토지피복도를 제공하고 있다.
- 토양도는 농촌진흥청의 흙토람을 통해 획득할 수 있으며, 토양의 각종 물리·화학적 조성에 대한 정보를 제공한다.
- 수치표고모델은 국토교통부의 국가공간정보포털을 통해 획득할 수 있으며, 5m 해상도의 수치표고모델을 활용할 계획이다.



<그림 2-10> 공간자료의 종류 및 예시

✓ **중권역 대표지점 목표 수질기준 현황**

- 중권역 대표지점은 물환경 목표기준 평가 규정(환경부고시 제 2018-6호)에 따라 생물화학적산소요구량(BOD), 총인(T-P) 항목에 대해 각각 평가하고 있으며, 수질측정 보고자료의 연간산술평균값을 활용하여 평가한다.
- 이때, 각 항목의 목표기준은 하천의 생활환경기준을 따르며, 중권역 대표지점별로 매우좋음(Ia)에서 약간나쁨(IV) 사이의 목표기준을 지니고 있다 <표 2-2>.

<표 2-2> 하천의 생활환경기준 (환경정책기본법시행령)

등급	기준								
	수소이온농도(pH)	생물화학적 산소요구량(BOD)(mg/L)	화학적 산소요구량(COD)(mg/L)	총유기탄소량(TOC)(mg/L)	부유물질량(mg/L)	용존산소량(mg/L)	총인(T-P)(mg/L)	대장균군(군수/100mL)	
								총대장균군	분원성대장균군
매우좋음(Ia)	6.5~8.5	1 이하	2 이하	2 이하	25 이하	7.5 이상	0.02 이하	50 이하	10 이하
좋음(Ib)	6.5~8.5	2 이하	4 이하	3 이하	25 이하	5.0 이상	0.04 이하	500 이하	100 이하
약간좋음(II)	6.5~8.5	3 이하	5 이하	4 이하	25 이하	5.0 이상	0.1 이하	1,000 이하	200 이하
보통(III)	6.5~8.5	5 이하	7 이하	5 이하	25 이하	5.0 이상	0.2 이하	5,000 이하	1,000 이하
약간나쁨(IV)	6.0~8.5	8 이하	9 이하	6 이하	100 이하	2.0 이상	0.3 이하	-	-
나쁨(V)	6.0~8.5	10 이하	11 이하	8 이하	쓰레기 등이 떠있지 아니할 것	2.0 이상	0.5 이하	-	-
매우나쁨(VI)	-	10 초과	11 초과	8 초과	-	2.0 미만	0.5 초과	-	-

- 2021년 물환경측정망 설치운영계획에 따르면 한강 대권역의 중권역 대표지점들의 경우 목표기준이 매우좋음(Ia)인 지점은 남한강 수계 4개소, 북한강 수계 7개소, 한강수계 1개소, 한강동해 수계 2개소로 총 14개 지점이며, 좋음(Ib)인 지점은 남한강 수계 3개소, 한강 수계 5개소, 한강동해 수계 1개소로 총 9개 지점이며, 약간좋음(II)은 남한강 수계 1개소, 한강서해 1개소로 총 2개 지점, 보통(III)은 한강수계 3개 지점, 약간나쁨(IV)는 한강서해 1개 지점이다.
- 금강 대권역의 중권역 대표지점들의 경우 목표 기준이 매우좋음(Ia)인 지점은 금강 수계 6개소, 만경·동진 수계 1개소로 총 7개 지점이며, 좋음(Ib)인 지점은 금강 수계 4개소, 금강서해 수계 1개소로 총 5개 지점이며, 약간좋음(II)은 금강 수계 1개소, 금강서해 2개소, 망경·동진 1개소로 총 4개 지점, 보통(III)은 금강 수계 1개 지점 만경동진 수계 1개소로 총 2개 지점이며, 약간나쁨(IV)는 삽교천 1개 지점이다.
- 낙동강 대권역의 중권역 대표지점들의 경우 목표 기준이 매우좋음(Ia)인 지점은 낙동강 수계 8개소, 낙동강동해 수계 2개소로 총 10개 지점이며, 좋음(Ib)인 지점은 낙동강 수계 12개소, 낙동강남해 수계 3개소, 낙동강동해 수계 1개소, 회야·수영 수계 1개소로 총 17개 지점이며, 약간좋음(II)은 낙동강 수계 2개소, 태화강 1개소, 형산강 1개소로 총 4개 지점, 보통(III)은 회야·수영 1개 지점이다.
- 영산·섬진강 대권역의 중권역 대표지점들의 경우 목표 기준이 매우좋음(Ia)인 지점은 섬진강 수계 2개소, 섬진강 남해 수계 2개소, 제주도 4개소로 총 8개 지점이며, 좋음(Ib)인 지점은 섬진강 수계 7개소, 섬진강 남해 수계 5개소, 영산강 수계 3개소, 영산강남해 수계 1개소, 영산강 서해 수계 2개소, 탐진강 1개소로 총 19개 지점이며, 약간좋음(II)은 영산강 수계 3개소, 보통(III)은 영산강 수계 2개소, 영산강 서해 1개소로 총 3개 지점이다.

<표 2-3> 한강 대권역 중권역 대표지점 별 수질목표 (환경정책기본법시행령)

대권역	수질목표	중권역 대표 지점
한강	매우좋음(Ia)	남한강: 4개 지점(남한강상류, 평창강, 충주댐, 충주댐하류) 북한강: 7개 지점(평화의댐, 춘천댐, 인북천, 소양강, 의암댐, 홍천강, 청평댐) 한강: 1개 지점(팔당댐) 한강동해: 2개 지점(양양남대천, 삼척오십천)
	좋음(Ib)	남한강: 3개 지점(달천, 섬강, 남한강하류) 한강: 5개 지점(한강잠실, 한강서울, 임진강상류, 한탄강, 임진강하류)

<표 2-3> 이어서

한강	좋음(Ib)	한강동해: 1개 지점(강릉남대천)
	약간좋음(II)	남한강: 1개 지점(경안천) 한강서해: 1개 지점(한강서해)
	보통(III)	한강: 3개 지점(한강고양, 한강하류, 안성천)
	약간나쁨(IV)	한강서해: 1개 지점(시화호)

<표 2-4> 금강 대권역 중권역 대표지점 별 수질목표 (환경정책기본법시행령)

대권역	수질목표	중권역 대표 지점
금강	매우좋음(Ia)	금강: 6개 지점(용담댐하류, 무주남대천, 영동천, 초강, 대청댐상류, 대청댐) 만경동진: 1개 지점(새만금)
	좋음(Ib)	금강: 4개 지점(용담댐, 보청천, 대청댐하류, 논산천) 금강서해: 1개 지점(금강서해)
	약간좋음(II)	금강: 1개 지점(미호천, 금강공주, 금강하구언) 금강서해: 2개 지점(대호방조제, 부남방조제) 만경・동진: 1개 지점(동진강)
	보통(III)	금강: 1개 지점(갑천) 만경동진: 1개 지점(만경강)
	약간나쁨(IV)	삽교천: 1개 지점(삽교천)

<표 2-5> 낙동강 대권역 중권역 대표지점 별 수질목표 (환경정책기본법시행령)

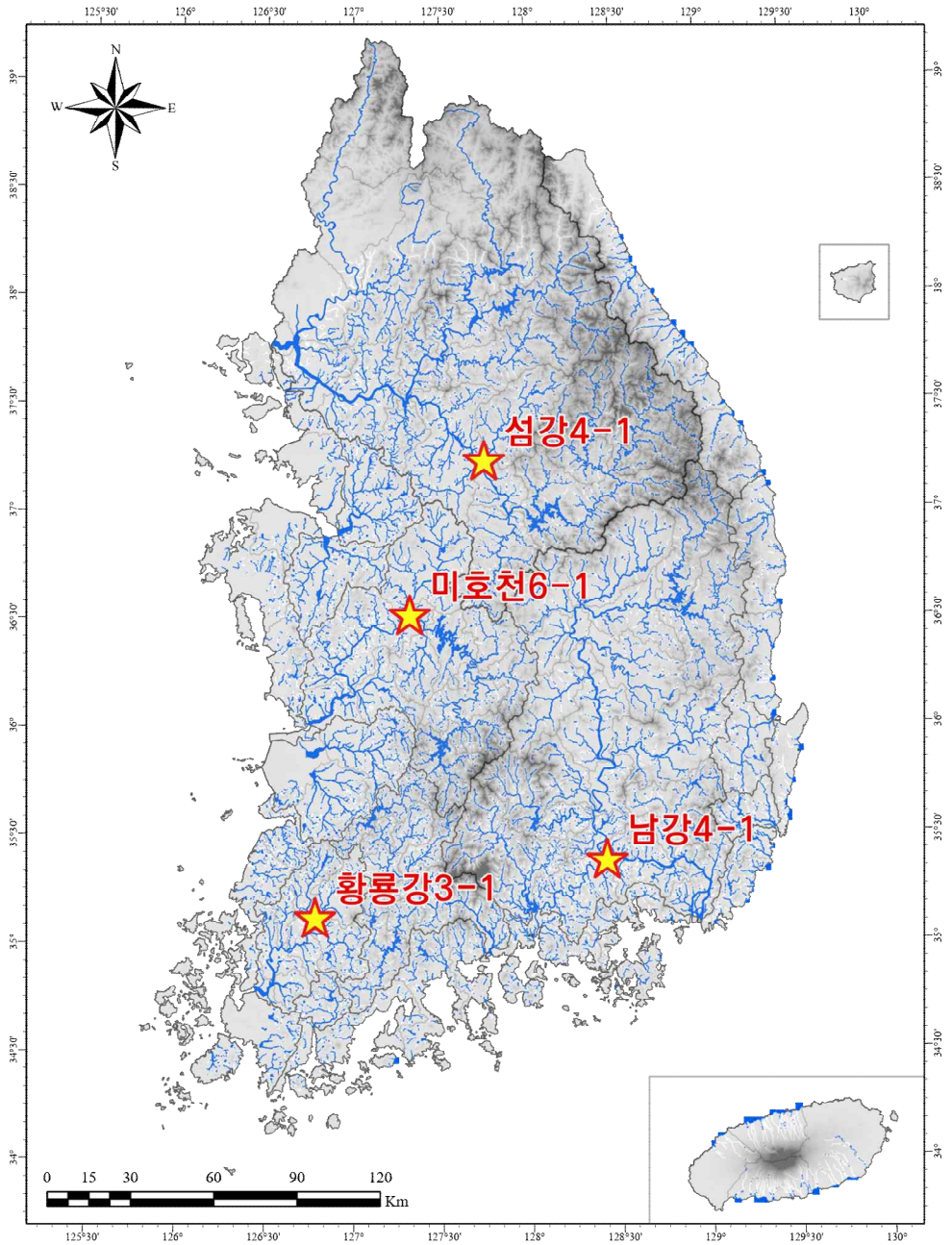
대권역	수질목표	중권역 대표 지점
낙동강	매우좋음(Ia)	낙동강: 8개 지점(안동댐, 안동댐하류, 내성천, 영강, 낙동상주, 구미보, 감천, 황강) 낙동강동해: 2개 지점(왕피천, 영덕오십천)
	좋음(Ib)	낙동강: 12개 지점(임하댐, 병성천, 위천, 강정고령보, 회천, 합천댐, 낙동창녕, 남강댐, 남강, 낙동밀양, 밀양강, 낙동강하구언) 낙동강남해: 3개 지점(가화천, 거제도, 낙동강남해) 낙동강동해: 1개 지점(대종천) 회야・수영: 1개 지점(수영강)
	약간좋음(II)	낙동강: 2개 지점(금호강, 창녕합천보) 태화강: 1개 지점(태화강) 형산강: 1개 지점(형산강)
	보통(III)	회야・수영: 1개 지점(회야강)

<표 2-6> 영산·섬진강 대권역 중권역 대표지점 별 수질목표 (환경정책기본법시행령)

대권역	수질목표	중권역 대표 지점
영산·섬진강	매우좋음(Ia)	섬진강: 2개 지점(주암댐, 보성강) 섬진강남해: 2개 지점(이사천, 수어천) 제주도: 4개 지점(제주서해, 제주북해, 제주남해, 제주동해)
	좋음(Ib)	섬진강: 7개 지점(섬진강댐, 섬진강댐하류, 오수천, 순창, 요천, 섬진곡성, 섬진강하류) 섬진강남해: 5개 지점(섬진강서남해, 완도, 금산면, 여수시, 남해도) 영산강: 3개 지점(영산강하류, 영암천, 영산강하구언) 영산강남해: 1개 지점(진도) 영산강서해: 2개 지점(와탄천, 신안군) 탐진강: 1개 지점(탐진강)
	약간좋음(II)	영산강: 3개 지점(황룡강, 지석천, 고막원천)
	보통(III)	영산강: 2개 지점(영산강상류, 영산강중류) 영산강서해: 1개 지점(주진천)

2-2-1-2. 연구 대상지역의 선정

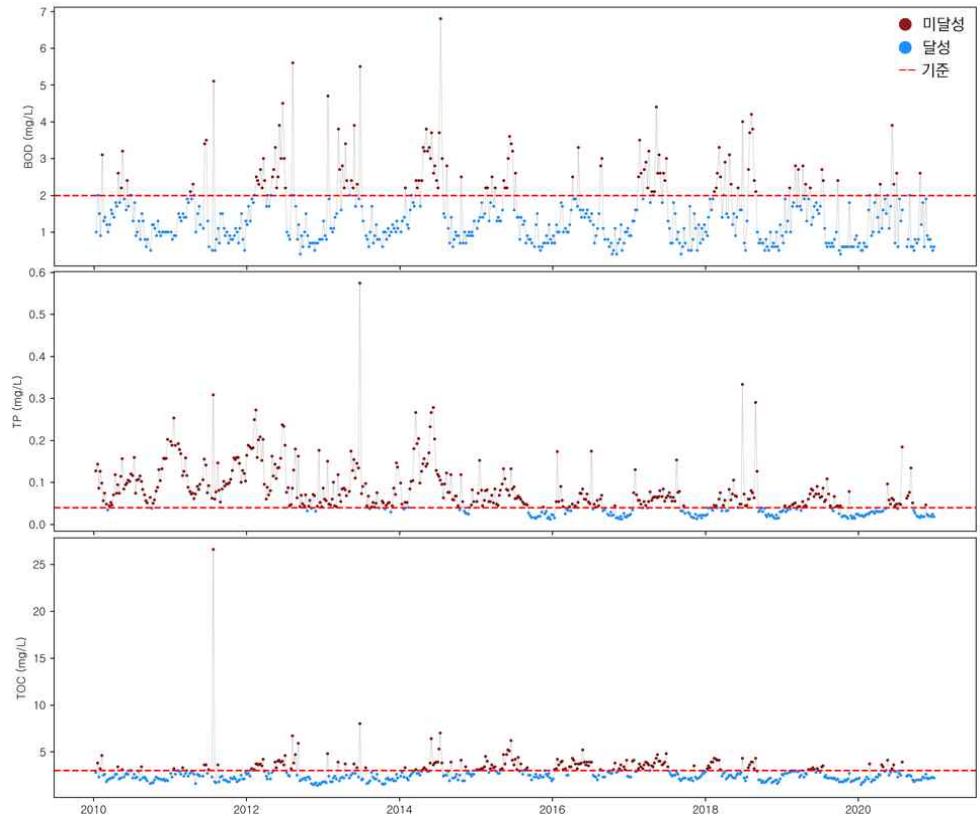
- 연구 대상지역은 가용 모니터링 자료의 극대화를 위해 중권역 대표지점 중 주요 지점인 후보 지류 수질측정망 지점 중 지점의 대표성, 발주처의 피드백 등을 종합적으로 고려하여 선정했다.
- 한강, 금강, 영산·섬진강, 낙동강의 각 대권역별로 각 1개 지점을 선정하였으며, 선정된 연구대상 지점은 한강 대권역 섬강4-1 지점, 금강 미호천6-1 지점, 영산·섬진강 황룡강 3-1 지점, 낙동강 대권역의 남강4-1지점이다.



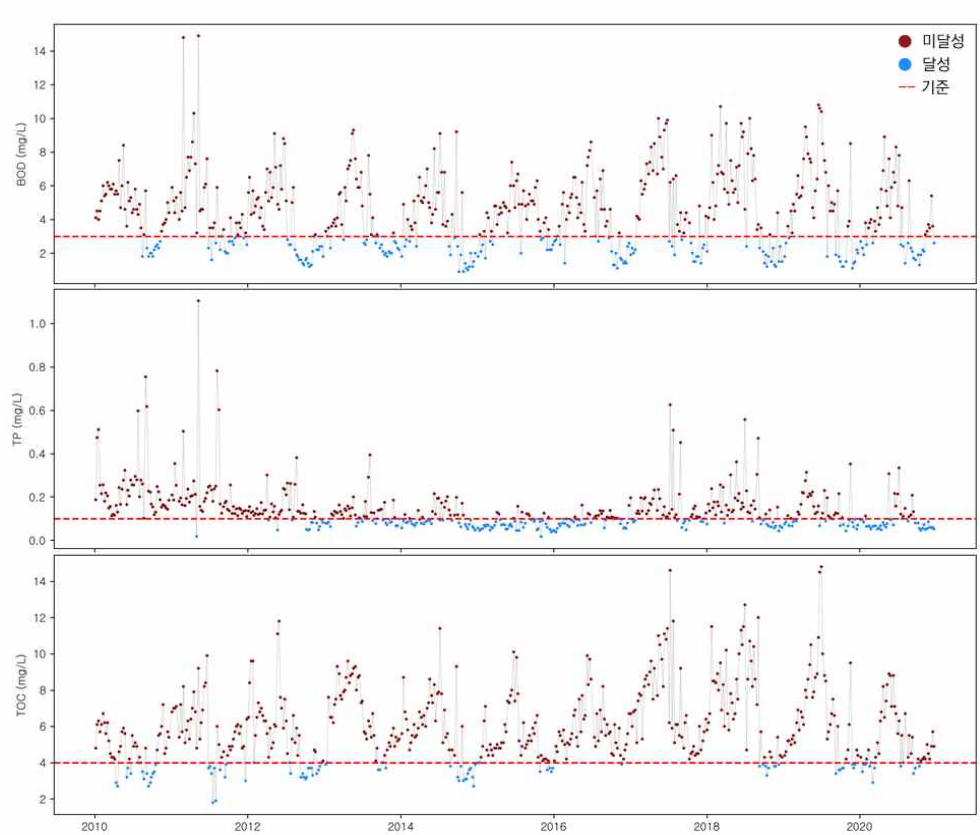
<그림 2-11> 연구 대상 지역

2-2-1-3. 연구 대상지역 현황에 대한 기초자료분석

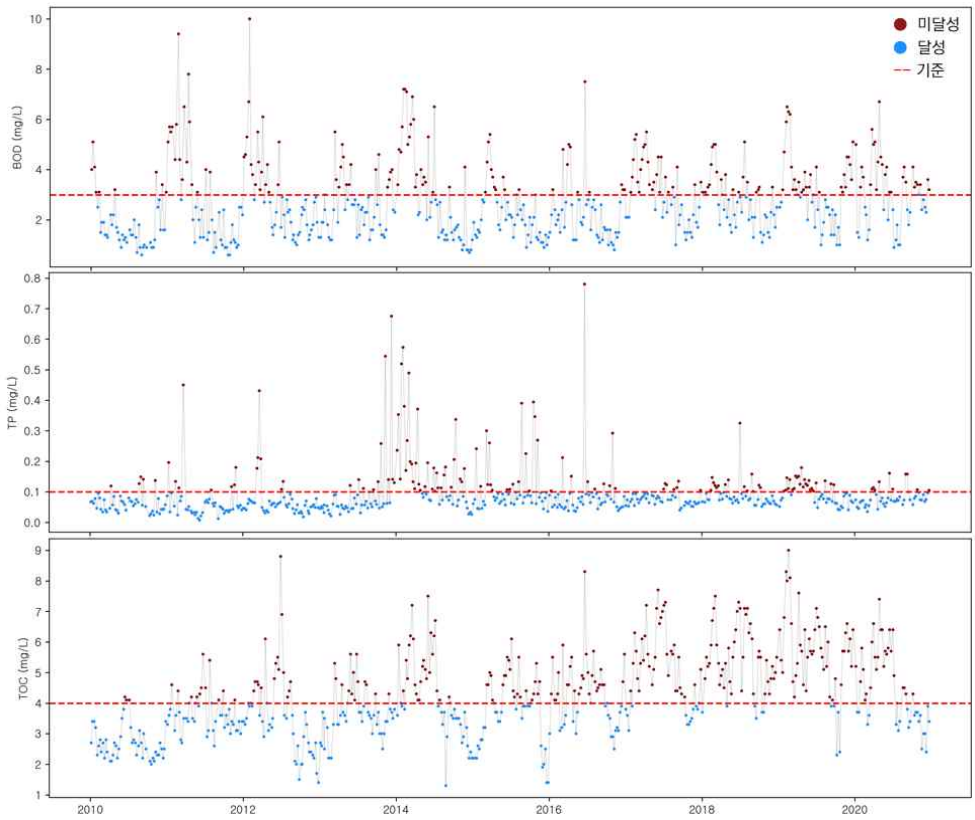
- 연구 대상지역의 수질 현황 파악을 위해 기초자료분석을 수행하였으며, 기초자료분석은 하천의 BOD, T-P, TOC 농도에 대해 수행하였다.
  - 이때, 중권역 대표지점은 연간산술평균을 기준으로 목표수질기준의 달성 및 미달성을 구분하나, 기초자료분석을 위해 각 측정일시별 농도값을 기준 초과 여부를 바탕으로 달성과 미달성의 범주로 구분하였다.
  - 기초자료분석을 위한 모니터링 자료 사용기간은 2010년부터 2020년까지이다.
- 한강 대권역의 연구 대상지역인 섬강4-1의 경우 목표기준이 좋음(Ib) 등급으로 전체 모니터링 결과에 대해 BOD기준(2mg/L 이하)을 21.96% 미달성했으며, T-P 기준(0.04mg/L 이하)은 71.25%, TOC 기준(3mg/L 이하)은 30.71% 미달성했다 <그림 2-12>.
- 금강 대권역의 연구 대상지역인 미호천6-1의 경우 목표기준이 약간 좋음(II) 등급으로 전체 모니터링 결과에 대해 BOD기준(3mg/L 이하)을 66.19%를 미달성했으며, T-P 기준(0.1mg/L 이하)은 55.93%, TOC 기준(4mg/L 이하)은 82.29%를 미달성했다 <그림 2-13>.
- 영산·섬진강 대권역의 연구 대상지역인 황룡강3-1의 경우 목표기준이 약간 좋음(II) 등급으로 전체 모니터링 결과에 대해 BOD기준(3mg/L 이하)을 36.87% 미달성했으며, T-P 기준(0.1mg/L 이하)은 26.26%, TOC 기준(4mg/L 이하)은 50.36% 미달성했다 <그림 2-14>.
- 낙동강 대권역의 연구 대상지역인 남강4-1의 경우 목표기준이 좋음(Ib) 등급으로 전체 모니터링 결과에 대해 BOD기준(2mg/L 이하)을 50.76% 미달성했으며, T-P 기준(0.04mg/L 이하)은 58.90%, TOC 기준(3mg/L 이하)은 61.36% 미달성했다 <그림 2-15>.



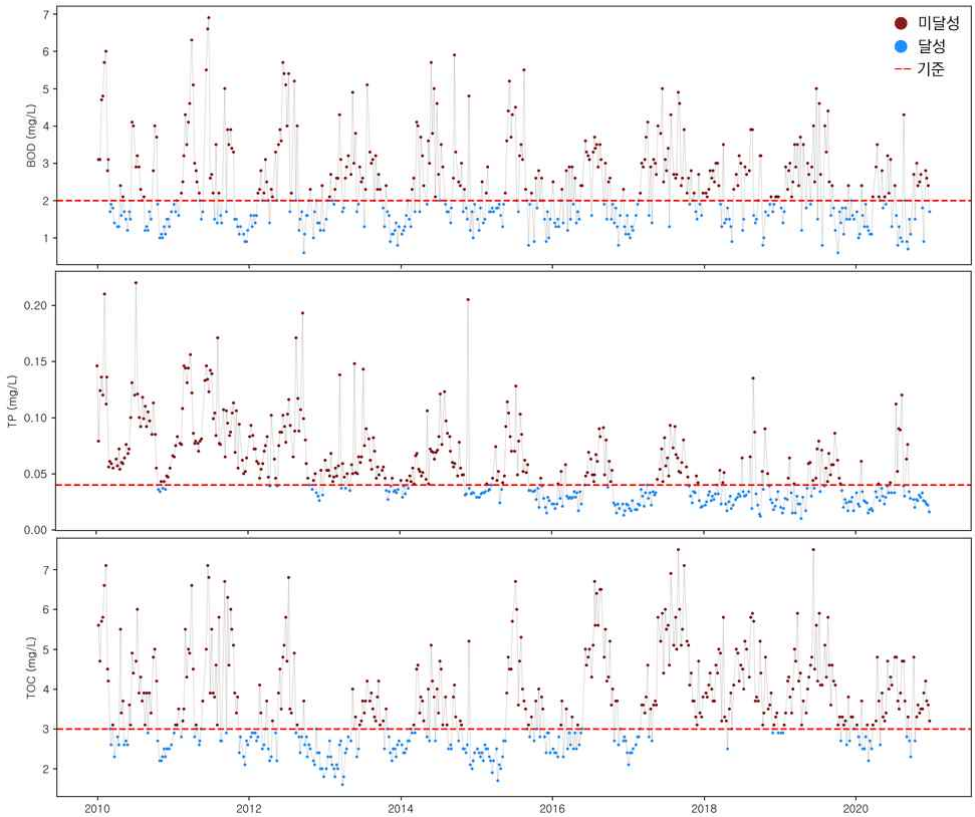
<그림 2-12> 2010-2020 기간 섬강4-1 지점 BOD, TOC, T-P 농도 변화



<그림 2-13> 2010-2020 기간 미호천6-1 지점 BOD, TOC, T-P 농도 변화



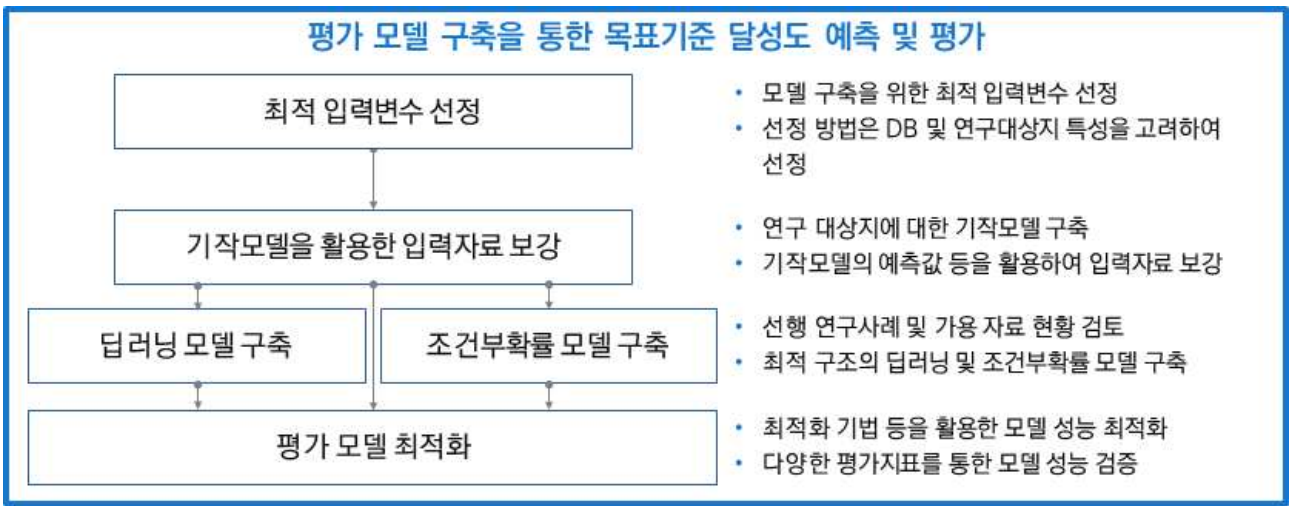
<그림 2-14> 2010-2020 기간 황룡강3-1 지점 BOD, TOC, T-P 농도 변화



<그림 2-15> 2010-2020 기간 남강4-1 지점 BOD, TOC, T-P 농도 변화

2-2-2 평가 모델 구축을 통한 목표기준 달성도 예측 및 평가

- 구축된 DB를 활용하여 연구대상지역의 목표 수질항목(BOD, T-P, TOC 등)의 중권역 수질목표 달성 여부 등에 대한 분석을 위한 딥러닝 및 조건부확률 모델을 구축하는 단계로 최적 입력변수 선정, 기작모델을 활용한 입력자료 보강, 딥러닝 및 조건부확률 모델 구축, 구축된 평가 모델 최적화로 구성한다 <그림 2-16>.
  - 최적 입력변수 선정 단계에서는 구축된 DB의 현황 및 연구 대상지의 특성 등을 검토한 뒤, 단계적 회귀분석, 요인분석 등을 활용하여 최적 입력변수를 선정하는 단계로 입력변수 선정을 위한 방법론은 연구 대상지역의 자료 형태 등을 고려하여 최적 기법을 유동적으로 선정하여 수행한다.
  - 기작모델을 활용한 입력자료 보강 단계는 연구 대상지역에 대해 SWAT 등 기작모델을 구축하고 모델을 통해 얻어진 예측값을 활용, 입력자료를 보강하는 단계이다.
  - 딥러닝 모델 및 조건부확률 모델 구축단계는 구축된 DB와 보강된 자료 등을 활용해 연구 대상지 및 분석 목적에 맞는 최적 모델을 구축, 학습하는 단계이다.
  - 이때, 딥러닝 모델 및 조건부확률 모델의 구축은 선행 연구사례 및 가용 자료 현황 등을 종합적으로 고려하여 구조 등을 선정함으로써 진행한다.
  - 이후, 최적화 기법 등을 활용하여 모델의 하이퍼파라미터를 최적화하고 우수한 예측 성능 확보를 도모하며, 다양한 평가지표를 활용하여 모델의 성능을 평가한다.



<그림 2-16> 평가모델 구축을 통한 목표기준 달성도 평가 흐름도 (안)

2-2-2-1. 최적 입력변수 선정

- 입력변수의 선정은 연구 진행에 있어 필요한 경우 진행하도록 한다.

✓ **단계적 회귀분석(Stepwise regression)**

- 단계적 회귀분석은 독립변수들 중 종속변수에 큰 영향을 미치는 독립변수를 식별하기 위해 사용되는 통계 분석도구이다.
- AIC(Akaike Information Criterion)이나 VIF(Variance Inflation Factor)과 같은 지표를 바탕으로 가장 유의하지 않은 독립변수를 회귀식에서 제외하거나 추가하면서 최적 회귀식을 구축한다.
- AIC는 회귀식을 과적합하여 너무 많은 독립변수를 포함시키는 것에 대한 패널티 함수로, 값이 작을수록 최적 모형으로 판단한다.

$$AIC = 2k - 2\ln(\hat{L}) \quad (1)$$

k : 변수 개수,

$\hat{L}$: 모델 최대 우도값

- VIF는 독립변수간의 다중공선성을 나타내는 지표로 다중공선성이 클수록 VIF 값이 커지게 된다.

$$VIF_i = \frac{\sigma^2}{(n-1)\,Var|X_i|} \cdot \frac{1}{1-R_i^2} \quad (2)$$

VIF_i : 회귀식의 i 번째 변수의 VIF

σ^2 : MSE, R_i^2 : 결정계수

- 다중공선성이란 회귀분석에서 독립변수들 간에 강한 상관관계를 나타내는 것을 의미하며, 다중공선성이 큰 경우 회귀분석의 전제 가정을 위반하게 된다.
- 따라서, 단계적 회귀분석을 선정하여 변수를 선택하는 경우, 단계적 회귀분석을 통해 얻어진 독립변수에 대해 VIF를 검토, 기준 (e.g. VIF < 10)을 넘지 않는 독립변수만을 선정하여 모델의 독립변수로 활용한다.

✓ **주성분 분석(Principal component analysis)**

- 주성분 분석은 각 대상 항목들의 선형 결합(linear combination)의 관계로 이루어진 새로운 축, 주성분을 생성함으로써 데이터 내에서 가장 유의미한 변수들을 식별하는 기법이다.
- 주성분 축의 방향은 자료의 최대분산을 설명하는 방향으로 결정되며 그에 따라 주성분 분석을 통해 데이터가 지닌 정보의 손실을 최소화 하면서 전체 데이터에 대한 분산을 가장 잘 표현할 수 있는 유의미한 변수들을 식별할 수 있다.

$$y_{i,j} = w_{1i}x_{1i} + w_{2i}x_{2i} + \ldots + w_{\pi}x_{pj} \quad (3)$$

y : 주성분 점수 (principal component score)
w : 변수 x와 y 간의 상관계수 (component loading)
i : 주성분의 수
j : 데이터의 수
p : 총 변수의 수

✓ **요인 분석(Factor analysis)**

- 요인 분석은 데이터 속의 잠재요인 혹은 구조를 알아내 그 구조를 쉽게 이해하고 해석할 수 있게 해주는 통계 방법이다.
- 전반적인 수질항목 간 상호작용을 주도하는 주요 수질항목을 선택하기 위해 데이터 변수들 사이에서 공통적으로 공유된 분산(공분산)을 설명할 수 있는 공통요인분석을 기반으로, 탐색적 요인분석(EFA, Exploratory Factor Analysis) 등이 있다.
- 요인 분석은 데이터 전처리, 요인 분석 적합도 검사, 요인 수효의 결정, 요인 추출 및 회전, 요인 분석 결과 및 해석의 총 6개의 단계로 진행한다.

- 데이터 전처리 단계에서는 환경 데이터에 알맞은 통상적인 이상치 제거 및 결측치 보완을 진행하고, 개별 항목의 정규화 과정을 통해 정규성 여부를 확인한다.
  - 이때 원래 데이터에서 정규성이 검증되지 않은 경우, 로그 변환, 지수 변환 등을 사용하여 데이터를 변환시켜준다.

- 요인 분석 적합도 검사 단계에서는 Kaiser Mayer-Olkin(KMO) 검정이나 Bartlett 검정과 같은 적합성 검정을 진행한다.

- 이때 KMO 검정 결과는 0에서 1까지의 값을 가지며, 최소 0.5 이상의 값을 가져야 요인 분석 진행이 가능하고 Bartlett 검정 결과 p-value가 0.05보다 작아야 데이터가 요인 분석에 적합하다고 판단한다.

- 요인 수효의 결정단계에서는 요인 분석을 통해 추출할 주요 수질항목의 개수를 판단한다.
  - 수효의 결정은 특정한 방식에 국한되지 않지만, 간단하고 널리 이용되는 Kaiser의 규칙을 사용하여 요인 수효 결정 방식이 권장된다.
  - Kaiser의 규칙은 데이터베이스의 대각행렬을 산정하고 존재하는 고윳값(Eigenvalue)을 산정하여 1보다 큰 고윳값의 개수를 요인의 수효로 정한다.

- 요인 추출 및 회전 단계는 데이터에서 요인의 특성을 추출하는 과정이다.
  - 대표적인 방법으로 주성분 분석(Principal component analysis, PCA)나 최우도법(Maximum likelihood estimation, MLE)이 있으며, 자료 형태 등에 따라 선택하여 활용이 가능하다.
  - 이때, 축의 회전에는 Varimax, Promax, Oblimin 등의 방법을 활용할 수 있으며, 추출된 축이 하나의 축에 분산의 설명이 집중되는 것을 방지하여 요인들의 개별적인 특징들을 분할시켜준다.

- 요인 분석 결과 도출 단계에서는 요인 분석결과로 얻어진 요인 수효와 동일한 개수의 요인을 도출하며, 이때 요인마다 각 수질 변수들의 요인적재량이 도출된다.
  - 요인적재량은 요인과 각 수질 변수들의 상관성 -1에서 1 사이의 값으로 정량화한 값으로 특정 요인에서 요인적재량이 큰 수질항목들은 상호 관계성이 큰 항목들로 데이터의 특성이 유사한 의미를 지니고 있다고 판단 가능하다.

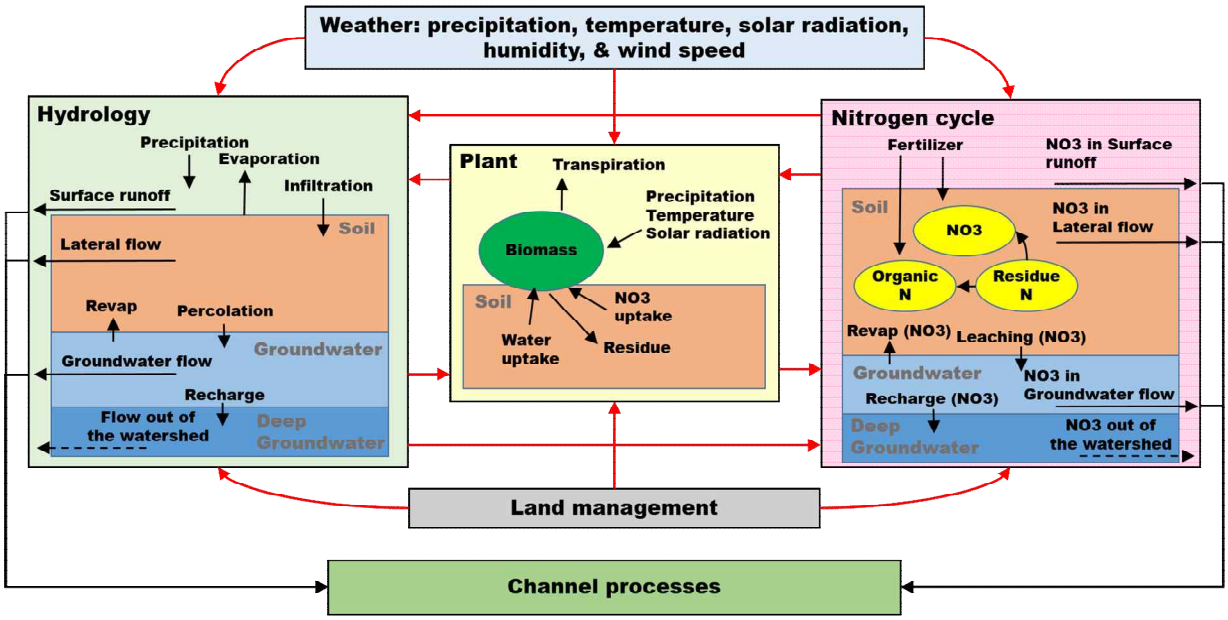
- 마지막으로 주요 수질 항목의 선정 단계에서는 각 요인들에서 요인적재량의 절대값이 가장 큰 항목을 그 요인의 특성을 대표하는 주요 특성으로 판단하고 그 항목을 그 요인의 특성을 대표하는 환경 변수로 판단한다.

2-2-2-2 . 기작모델을 활용한 입력자료 보강

- 월 단위로 수행되는 모니터링 자료의 특성상 시간적 규모의 확장을 위해 기작모델 등을 활용하여 추가 입력자료의 보강을 도모한다.
- SWAT(Soil and Water Assessment Tool) 모델 등을 활용하여 연구 대상지역에 대한 모델링을 수행, 딥러닝 및 조건부확률 모델 구축을 위한 추가 입력자료를 생산한다.

✓ SWAT

- SWAT 모델은 미국 농무성 농업연구소(United States Department of Agriculture - Agricultural Research Service, USDA-ARS)에서 개발된 물리적 기반의 준 분포형 강우-유출 모델이다 <그림 2-17>.
- 유역단위 수문 및 수질모의에 효율적인 수단으로서 알려진 SWAT 모델은 다양한 국가에서 물환경 관리 분야에 적용되고 있다.



<그림 2-17> SWAT 모형 모식도

- SWAT은 수문, 토양유실, 영양물질, 하도추적 부모형으로 구성되어 물질순환을 모의한다.
  - 우선 모델 구동을 위해 다양한 공간단위로 대상지가 구분된다.
  - 지표면 고도자료 기반으로 유역을 구분한 후 지형특성 반영하여 수 개의 소유역으로 구분되고 소유역 내 고유한 토지피복 및 토양, 지형경사에 따라 모델링 최소단위인 수문반응단위(HRU, Hydrologic Response Unit)로 나뉜다.

- 각 HRU에서 물수지식에 근거해 표면 유출량, 지하수로의 침투량, 증발산량을 산정한다.
- HRU에서 산정된 물수지양은 소유역단위로 통합하여 계산되며 소유역간의 위치에 따라 상류에서 하류 유역으로 배출되는 유량을 산정한다.
- 따라서, SWAT모형에서 모의되는 물수지는 아래 수식에 기초한다.

$$SW_t = SW_0 + \sum_{i=1}^{t}(R_{day} - Q_{surf} - E_a - W_{seep} - Q_{gw}) \quad (4)$$

- 여기에서 SW_t는 최종 토양수분량(mm), SW_0는 i일 동안의 초기토양수분량(mm), t는 시간(days), R_{day}는 i일 동안의 강수량(mm), Q_{surf}는 i일 동안의 표면유출량(mm), E_a는 i일 동안의 증발산량(mm), W_{seep}는 i일 동안의 침루량(mm), Q_{gw}는 i일 동안의 회귀수량(mm)을 나타낸다.
- SWAT모델은 지표면에서 강우가 지하로 침투하는지 표면으로 유출되는지 모의하기 위해 SCS 유출곡선법과 Green&Ampt 침투법의 두 가지 방법을 제공한다.
- 잠재증발산량 모의를 위해 Penman Monteith, Priestly Taylor, Hargreaves 3가지 방법이 적용가능하다.
- SWAT 모델의 검보정은 예측하려는 변수(유량, T-P, T-N 등)와 관련된 매게변수 통계적 분석을 통해 모델의 재현성을 검증한다. 통계적 분석에 사용되는 대표적인 목적함수는 Nash-Sutcliffe efficiency(NSE), RMSE-observations standard deviation ratio(RSR), Percent bias(PBIAS)이며 아래와 같다.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O_i})^2} \quad (5)$$

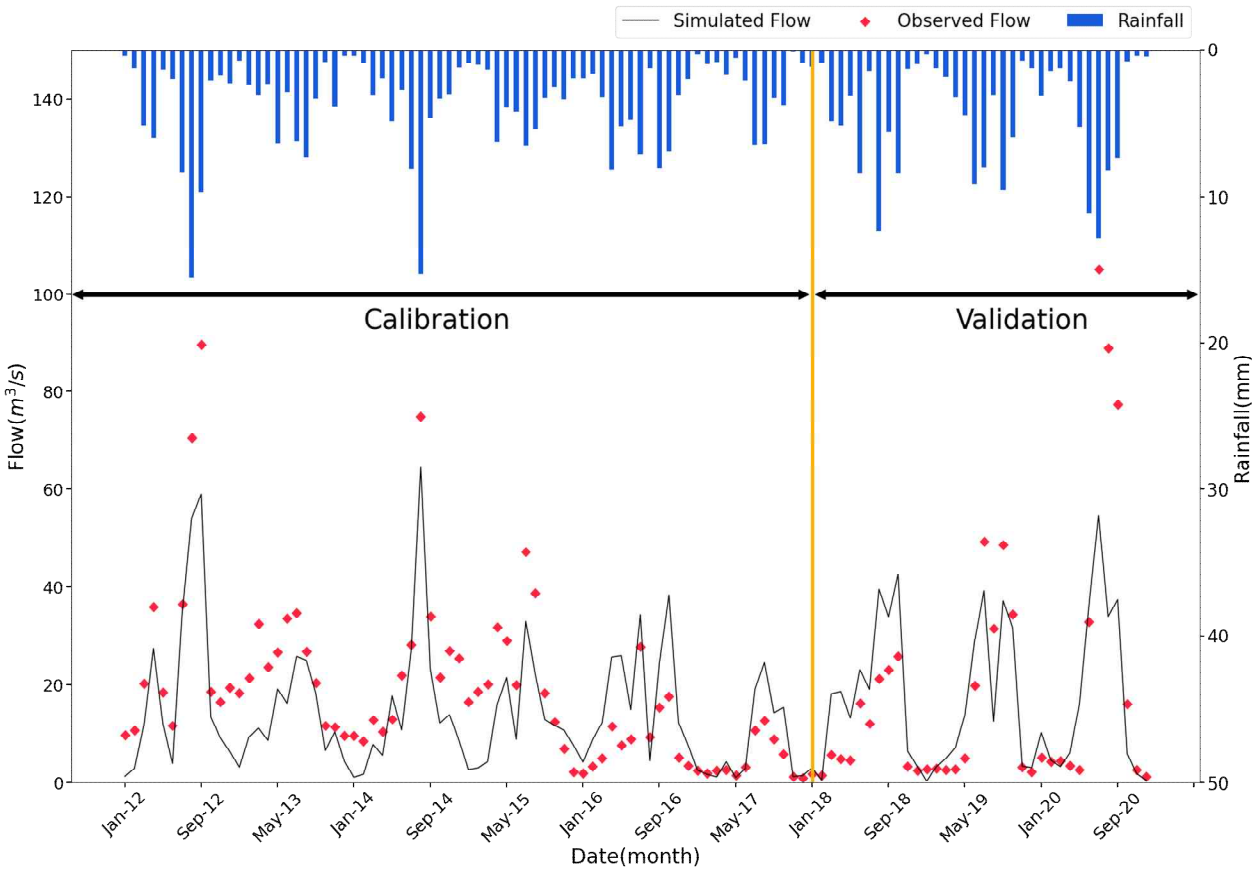
$$RSR = \frac{[\sqrt{\sum_{i=1}^{n}(O_i - P_i)^2}]}{[\sqrt{\sum_{i=1}^{n}(O_i - \overline{O_i})^2}]} \quad (6)$$

$$PBIAS = [\frac{\sum_{i=1}^{n}(O_i - P_i) \times 100}{\sum_{i=1}^{n}(O_i)}] \quad (7)$$

- 여기서 O_i는 각 강우시 실측된 값이고, P_i는 모델에서 예측된 각 강우별 모의 값이며, $\overline{O_i}$는 모든 강우시 실측값의 평균이다. NSE 값이 1에 가까울수록, RSR과 PBIAS 값이 0에 가까울수록 모델이 실측치를 잘 모사한 것이다.
- 월 단위 모의 결과에 따라 평가된 각 통계치에 따라 모델의 재현성은 4단계로 구분된다 <표 2-8>.

<표 2-7> 각 통계치 별 모델의 재현성 구분

재현성	RSR	NSE	PBIAS(%)		
			유량	Sediments	N, P
Very good	0.00<RSR<0.50	0.75<NSE<1.00	PBIAS<10	PBIAS<±15	PBIAS<±25
Good	0.50<RSR<0.60	0.65<NSE<0.75	±10<PBIAS <±15	±15<PBIAS <±30	±25<PBIAS <±40
Satisfactory	0.60<RSR<0.70	0.50<NSE<0.65	±15<PBIAS <±25	±30<PBIAS < ±55	±40<PBIAS <±70
Unsatisfactory	RSR>0.70	NSE<0.50	PBIAS>±25	PBIAS>±55	PBIAS>±70

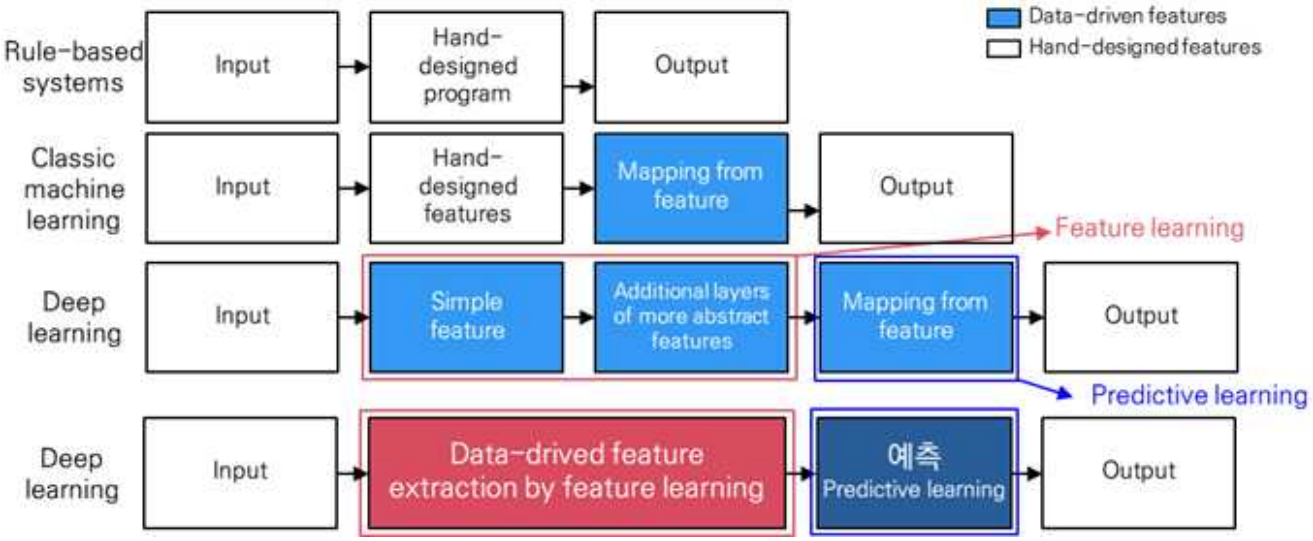


<그림 2-18> SWAT 모형을 활용한 유량 검보정 예시

2-2-2-3. 목표기준 예측을 위한 모델 구축 및 최적화

1-1). 딥러닝 모델 구축

- 딥러닝(Deep learning) 기법이란, 데이터의 특징을 추출하는 다양한 기계학습 알고리즘의 계층적 구조로 이루어진 알고리즘의 집단이다 <그림 2-19>.
- 따라서, 딥러닝 기법은 데이터를 기반으로 스스로 데이터에 내재 되어있는 중요한 정보를 추출(특징 추출)하고 이를 바탕으로 모델링의 목적에 따라 예측, 분류, 군집화 등의 작업을 수행할 수 있다.
- 최근 빅데이터 시대의 도래 등으로 인해 이미지 분석, 음성인식 등 다양한 분야에서 활용되고 있으며, 뛰어난 성능을 입증받아 왔다.

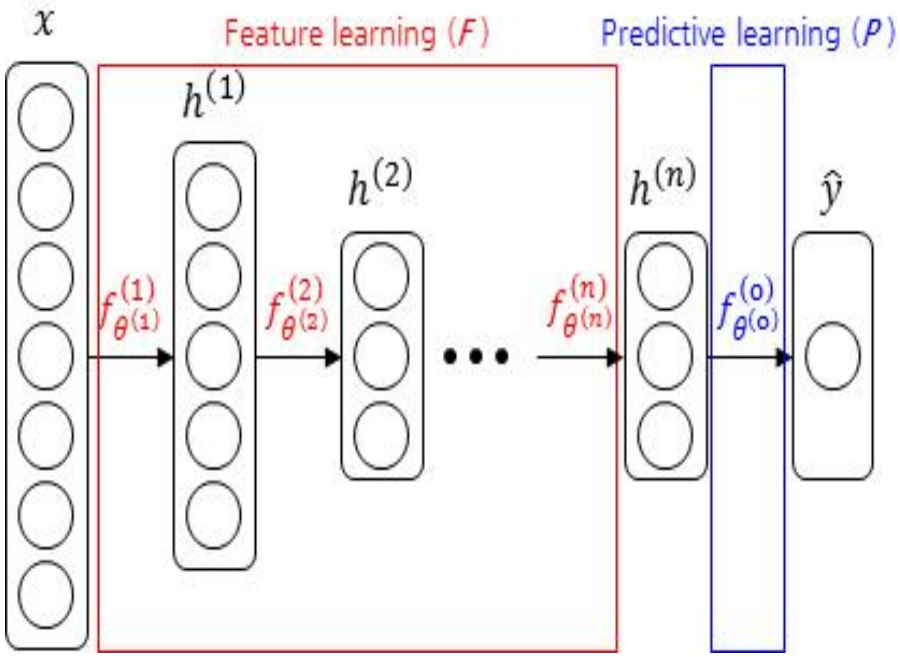


<그림 2-19> 딥러닝 모델의 특징, Goodfellow, 2016

✓ **심층신경망**

- 딥러닝의 기초가 되는 인공신경망은 뇌세포 내의 뉴런의 형태를 모방하여 입력값(input data)에 가중치(weight)를 가하여 정보를 전달해 나가는 구조로 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)의 세 개의 층으로 이루어져 있으며, 입력층에서 은닉층을 통하여 출력층까지 각 층(layer) 사이마다 존재하는 가중치를 순차적으로 갱신하며 예측값과 실제값과의 오차를 최소화하면서 학습을 수행한다.

- 일반적으로 여러 층(≥2)의 은닉층을 지닌 인공신경망을 심층신경망(deep neural network)라 하며 심층신경망의 구조는 아래와 같으며, 크게 특징학습(feature learning) 부분과 예측학습(predictive learning) 부분으로 나뉜다 <그림 2-20>.



<그림 2-20> 심층신경망의 구조

- 이때, 첫 번째 은닉 특징(hidden feature)은 초기 가중치와 입력값의 곱에 초기 bias를 더한 값에 활성화 함수(activation function)를 적용한 형태로 표현할 수 있으며, 아래와 같은 식으로 표현할 수 있다.

$$h^{(1)} = f^{(1)}_{\theta^{(1)}}(x) = f^{(1)}(x;\theta^{(1)}) = \sigma^{(1)}(W^{(1)}x + b^{(1)}) \quad (8)$$

$h^{(1)}$: 첫 번째 은닉특징
$f^{(1)}$: 첫 번째 은닉층
$\sigma^{(1)}$: 첫 번째 은닉층의 활성화 함수
$W^{(1)}$: 첫 번째 은닉층의 초기 가중치
$b^{(1)}$: 첫 번째 은닉층의 초기 bias
x : 입력 값
$\theta^{(1)}$: 첫 번째 은닉층의 가중치와 bias 등의 파라미터 집합

- 따라서, n개의 은닉층으로 이루어진 특징 학습 부분은 아래와 같이 표현할 수 있다.

$$h^{(n)} = F(x) = f^{(n)}_{\theta^{(n)}}(...f^{(k)}_{\theta^{(k)}}...(f^{(2)}_{\theta^{(2)}}(f^{(1)}_{\theta^{(1)}}(x))) \quad (9)$$

F(x) = 특징학습

$f^{(n)}$: n 번째 은닉층

x : $\theta^{(n)}$: n 번째 은닉층의 가중치와 bias 등의 파라미터 집합

- 특징학습 부분과 마찬가지로 예측을 수행하기 위한 예측학습 부분 또한 아래와 같이 표현할 수 있다.

$$\hat{y} = P(h^{(n)}) = f^{(o)}_{\theta^{(o)}}(h^{(n)}) = f^{(o)}(h^{(n)};\theta^{(o)}) = \sigma^{(o)}(W^{(o)}h^{(n)} + b^{(o)}) \quad (10)$$

$\hat{y}$: 모델의 예측값

P(x) : 예측학습 부분

$f^{(o)}$: 출력층

- 결과적으로, 심층신경망은 특징학습을 담당하는 은닉층과 예측학습을 담당하는 출력층의 결합이므로 n개의 은닉층으로 이루어진 DNN은 아래와 같이 표현할 수 있다.

$$\hat{y} = DNN(x) = P(F(x)) = f^{(o)}_{\theta^{(o)}}(f^{(n)}_{\theta^{(n)}}(...f^{(k)}_{\theta^{(k)}}...(f^{(2)}_{\theta^{(2)}}(f^{(1)}_{\theta^{(1)}}(x)))) \quad (11)$$

$\hat{y}$: 모델의 예측값, x : 입력 값

P(x) : 예측학습 부분, F(x) : 특징학습 부분

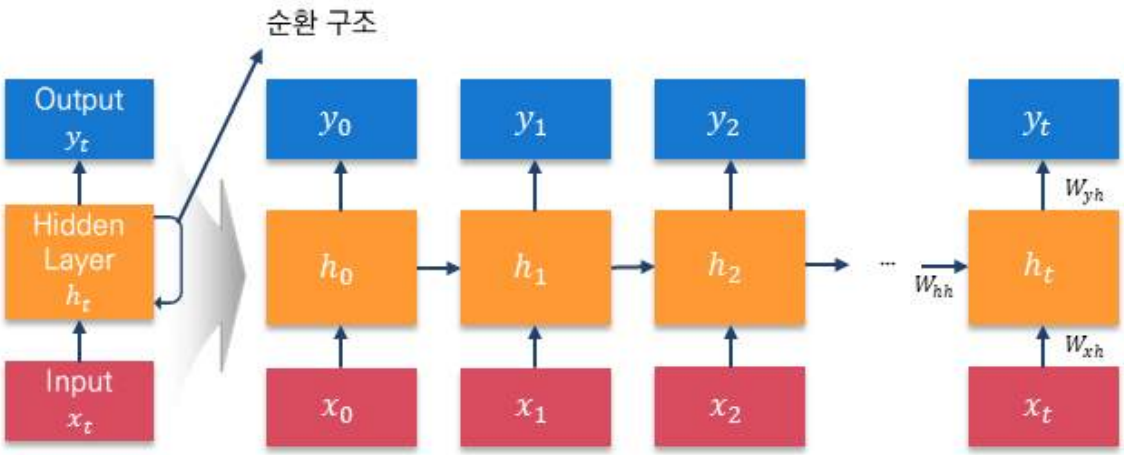
$f^{(n)}$: n 번째 은닉층

$\theta^{(n)}$: n 번째 은닉층의 가중치와 bias 등의 파라미터 집합

$f^{(o)}$: 출력층, $\theta^{(o)}$: 출력 층의 가중치와 bias 등의 파라미터 집합

✓ **순환신경망(Reccurrent Neural Network)**

- 기존의 인공신경망은 모든 입력과 출력이 독립적, 순차적으로 동작하여 데이터의 입력 데이터의 순서와 지속성 등을 고려할 수 없다.
- 순환신경망(Recurrent Neural Network, RNN)은 순환구조를 가진 인공신경망으로써, 순서가 있는 데이터는 시퀀스 데이터(sequence data)의 학습에 있어 입력 데이터의 순서를 반영하여 처리하는 구조를 지닌 신경망 모델이다.
- 순환신경망은 현재의 입력변수를 처리하기 위해 모델에 입력된 과거의 입력변수를 재귀(recursion)하여 현재와 과거의 독립된 입력변수를 연결하여 처리하며, 지난 시점의 은닉층을 현 시점(t)의 은닉층에 누적시켜 계산한다 <그림 2-21>.



<그림 2-21> 순환신경망의 구조

- 순환신경망은 이러한 특징으로 인해 이전의 계산결과가 다음 계산에 영향을 미치기 때문에 데이터의 순서가 정해져 있는 시퀀스 데이터, 특히 시계열 데이터를 처리하는데 적합하다고 알려져 있다.
- 정리하면, 순환신경망은 하나의 인공신경망이 여러 개 복사된 형태로 구성되어 있고 각 인공신경망은 시퀀스 데이터의 특정 시점 t에서 다른 시점인 t+1으로 은닉층 벡터를 넘겨주고 있는 형태로 시퀀스 데이터의 길이에 관계없이 자유롭게 입력변수와 출력변수 설정이 가능하다 <그림 2-22>.
- 이를 수식으로 표현하면 아래와 같으며, f_w는 활성화 함수, h_t는 t 시점의 은닉층으로 이는 t-1 시점의 은닉층과 매 시점마다 적용되는 입력값 벡터 x_t의 활성화 함수로 표현된다.

$$h_t = f_w(h_{t-1}, x_t) \tag{12}$$

f_w : 활성화 함수(Activation function)

h_t : 은닉층

h_{t-1} : 이전 시퀀스의 은닉층

x_t : 입력값

- 이때 순환신경망에서는 일반적으로 활용되는 활성화 함수인 하이퍼볼릭 탄젠트 함수를 사용하여 은닉층은 아래와 같이 표현할 수 있다.

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \tag{13}$$

W_{hh} : 이전 시점의 은닉층 벡터 h_{t-1}를 위한 가중치,

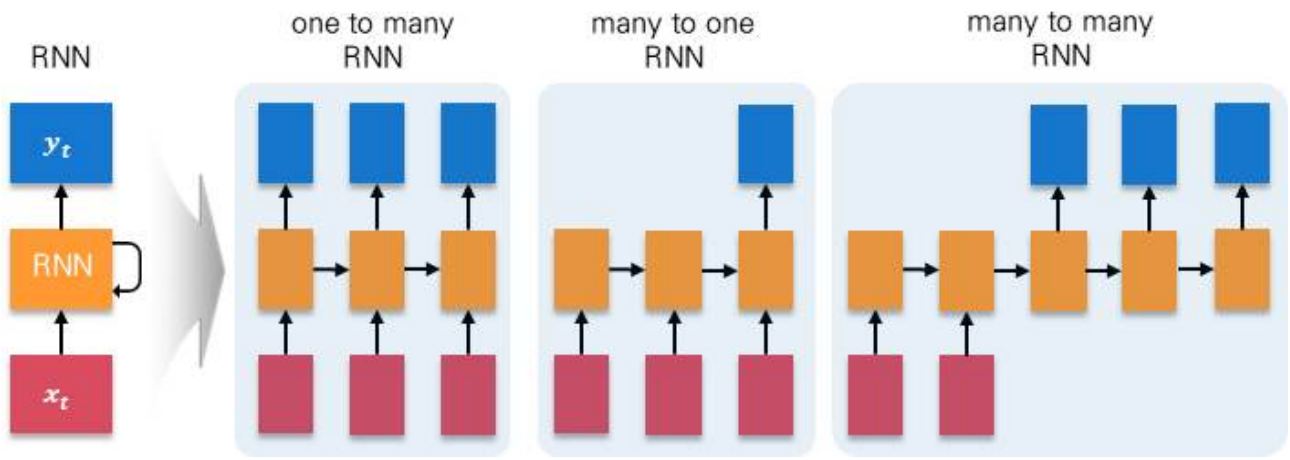
W_{xh} : 입력 벡터 x_t를 위한 가중치

- 이를 토대로 순환신경망의 출력값은 아래와 같이 표현할 수 있다.

$$y_t = W_{hy}h_t \tag{14}$$

y_t : 출력벡터

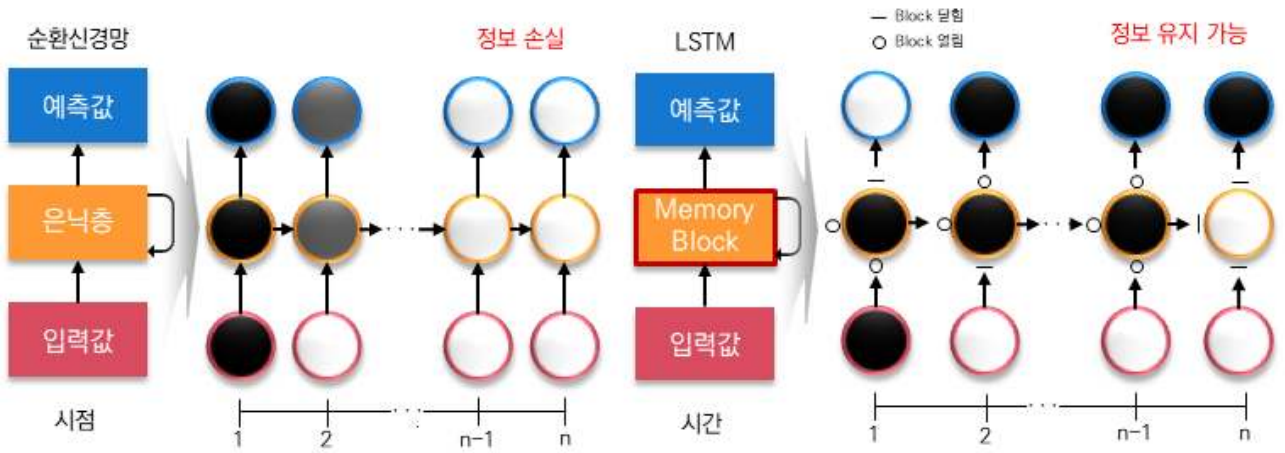
W_{yh} : y_t를 위한 가중치



<그림 2-22> 다양한 순환신경망의 구조 예시

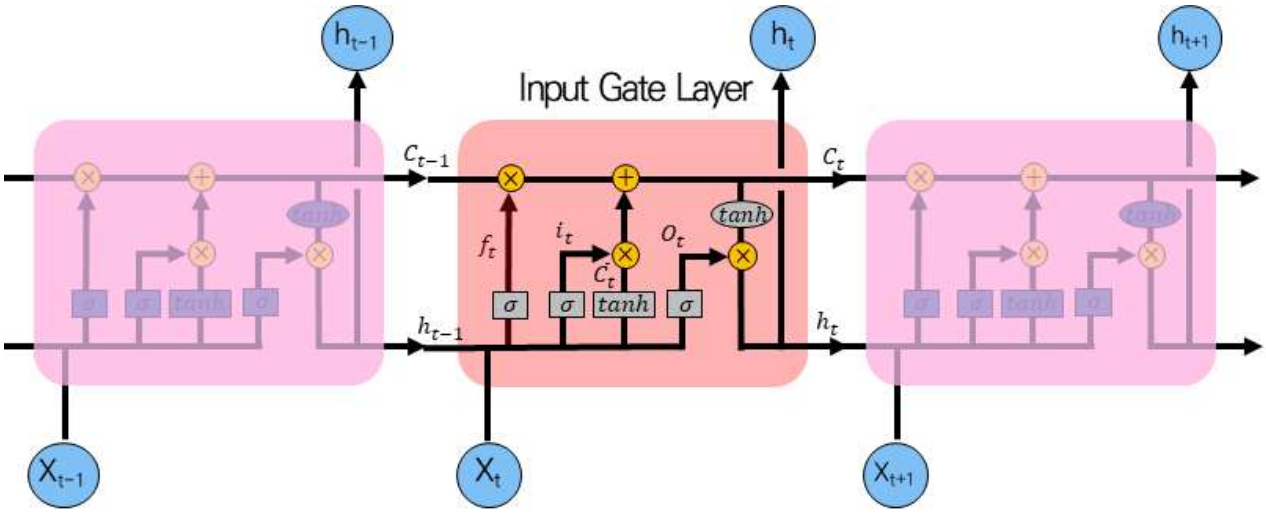
✓ 장단기 메모리(Long Short Term Memory, LSTM)

- LSTM은 순환형 구조로 인한 장기 의존성(long-term dependency)로 인한 과거 순서의 자료에 대한 가중치가 소실되는 가중치 소실(vanishing gradient) 문제를 해결하기 위해 제시되었다 <그림 2-23>.



<그림 2-23> LSTM의 장기의존성 및 가중치 소실 해결 방법

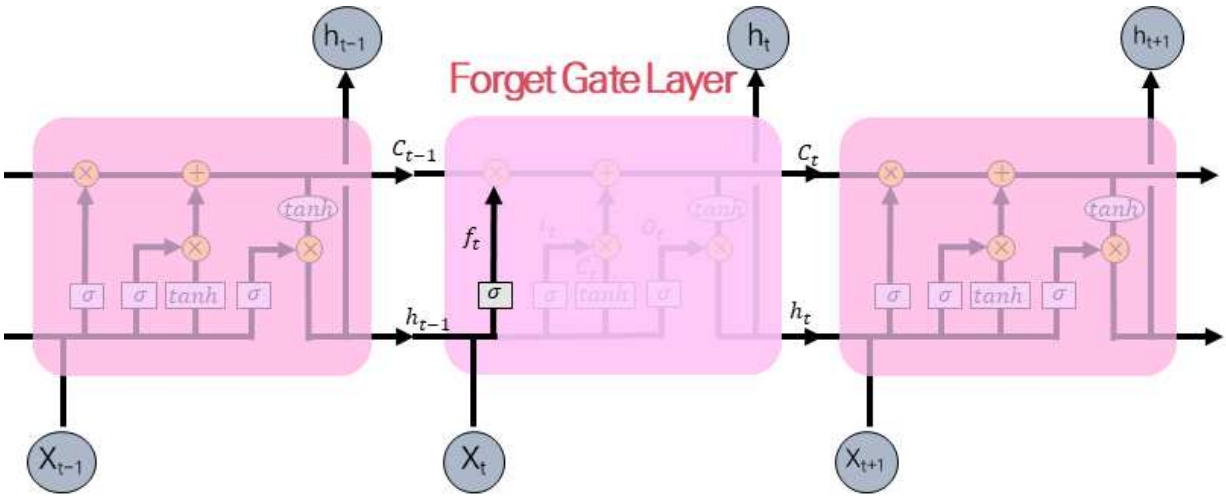
- LSTM은 가중치 반영 및 활성화 함수 변환을 통해 입력값 벡터에서 출력값 벡터로 변환하는 단계를 하나의 셀(cell)로 정의한다 <그림 2-24>.



<그림 2-24> LSTM cell의 구조

- 중단기 순환신경망은 셀 내부의 상태량인 셀 상태(cell state)를 망각, 입력, 갱신, 출력 총 4단계의 계산과정을 통해 가중치 소실, 장기의존성 문제가 발생하지 않도록 조절한다.
- LSTM의 작동 단계 중 첫 번째 단계는 망각단계로써, 특정 정보의 제거 여부를 망각 게이트(forget gate, f_t)를 활용하여 결정한다 <그림 2-25>.
  - 즉, 첫 번째 단계는 셀 상태(cell state)에서 어떤 정보를 버릴지, 유지할지를 결정하는 단계로써 이전 시점의 은닉층 벡터 h_{t-1}와 입력벡터 x_t를 받아 0과 1사이 값을 셀상태 C_{t-1}에 전송한다.
  - 이때 0은 정보의 완전 제거, 1은 정보의 완전 유지를 의미하며 중단기 순환신경망의 셀(cell)내에서 정보를 유지할지 제거할지 선택하게 된다(Jung et al., 2018).

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{15}$$



<그림 2-25> LSTM cell의 구조

- 두 번째 단계는 입력단계로써 입력게이트(input gate, i_t)를 통하여 새로운 정보의 저장여부를 결정하는 단계로 이후에 들어오는 새로운 정보 중 어떤 정보를 셀상태(cell state)에 저장할 것인지 결정하는 단계이다 <그림 2-26>.
  - 입력단계는 세부적으로 1) 입력게이트 층(Input gate layer)라 불리는 Sigmoid 활성화 함수 σ가 어떤 값을 업데이트 할지 정하는 단계; 2) 하이퍼볼릭탄젠트 함수 tanh가 셀상태에 더해질 수 있는 새로운 후보값의 벡터인 $\tilde{C}_t$를 생성하는 2단계로 구성된다.

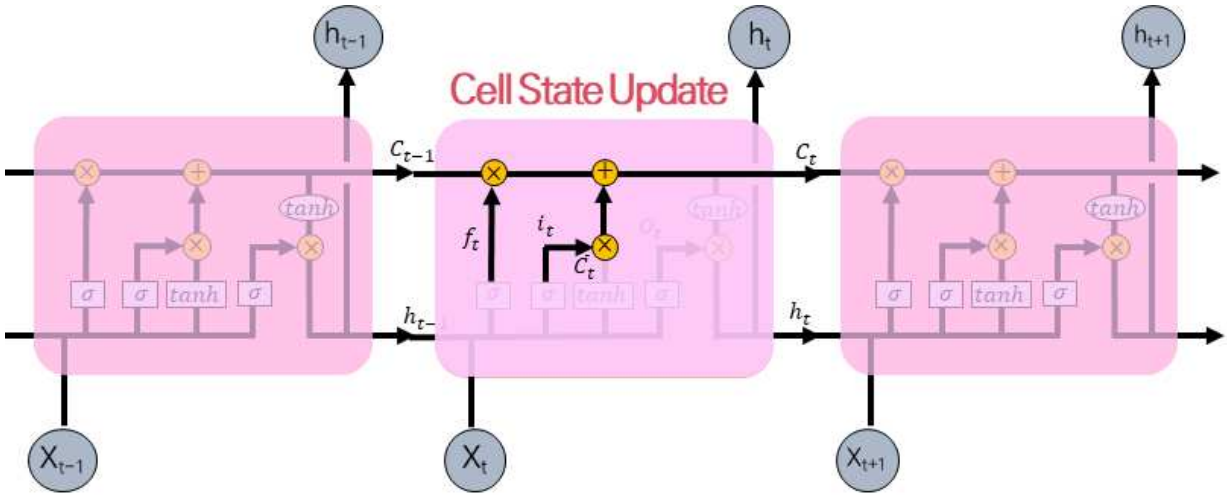
• 결과적으로 입력단계는 아래의 수식을 통해 최종 셀 상태를 갱신할 값을 만든다.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{16}$$

i_t : 0 또는 1의 값을 가지는 입력게이트 값

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_c) \tag{17}$$

$\tilde{C}_t$: tanh로 구성되어있는 셀 상태의 중간값



<그림 2-26> LSTM cell의 구조

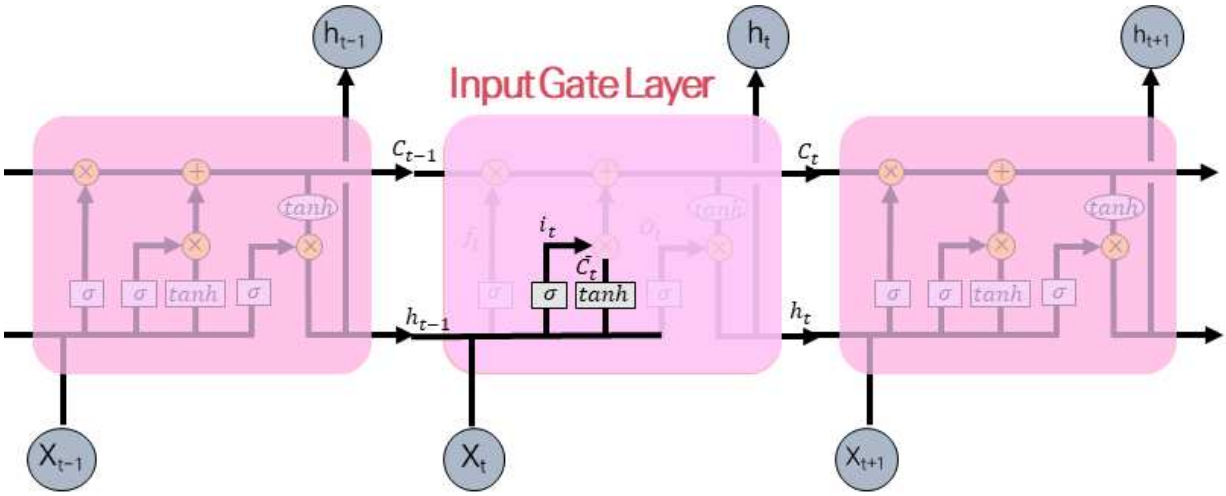
- 세 번째 단계는 갱신단계로써 입력게이트와 출력게이트의 값을 이용하여 셀 상태를 갱신한다.

• 갱신단계에서는 망각단계의 망각게이트 값 f_t를 과거 셀 상태값 C_{t-1}를 곱해서 망각단계에서 결정된 작업을 수행한 후 입력단계의 입력게이트 값 i_t과 셀 상태의 중간값 $\tilde{C}_t$의 곱을 통해 망각단계에서 갱신하기로 한 값을 얼마나 갱신할지 결정한다.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{18}$$

C_t : 현재 셀 상태값

C_{t-1} : 과거 셀 상태값



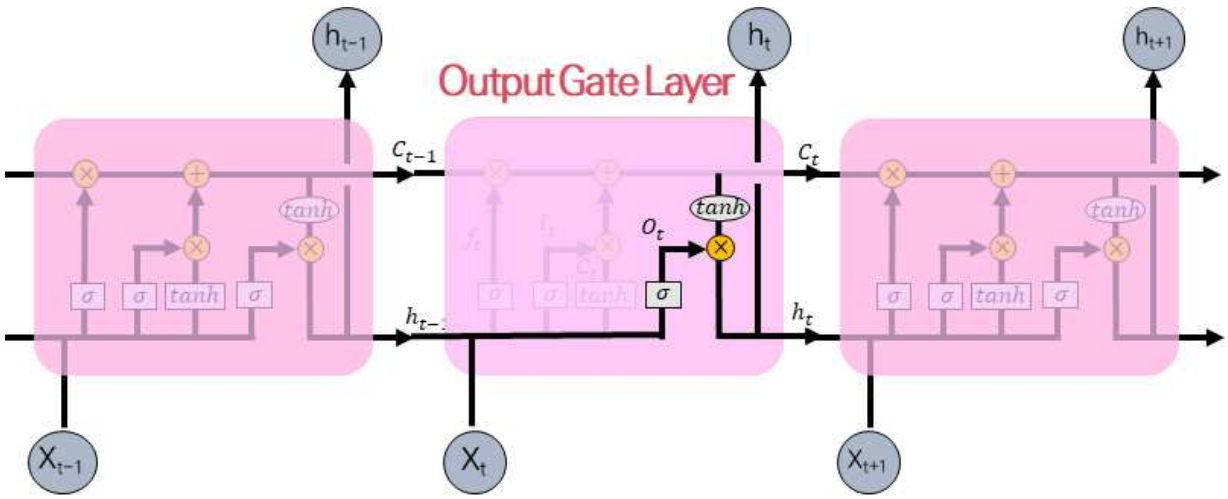
<그림 2-27> LSTM cell의 구조

- 마지막 단계는 출력단계로써 출력게이트와 셀 상태를 이용하여 출력값을 계산한다.
  - 출력단계에서는 tanh를 사용하여 망각, 입력, 갱신단계를 거쳐 활성화된 셀상태 C_t와 곱으로 특정 시점 t의 은닉층 벡터 h_t를 출력한다.

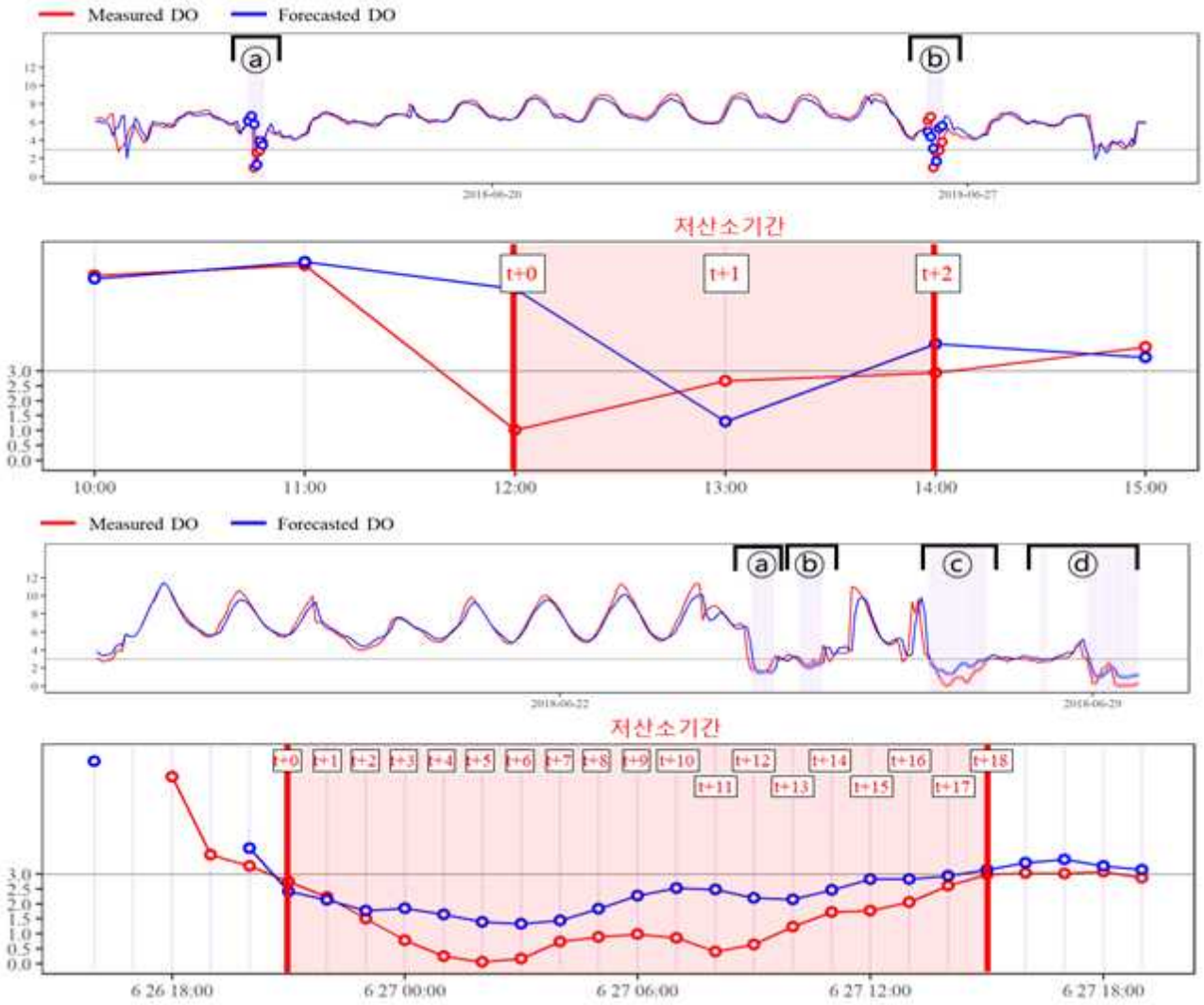
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{19}$$

$$h_t = o_t \bullet \tanh(C_t) \tag{20}$$

o_t : 0 또는 1의 값을 가지는 출력게이트 값



<그림 2-28> LSTM cell의 구조



<그림 2-29> LSTM을 활용한 저산소 시기 용존산소 농도 예측 예시

1-2). 딥러닝 모델의 하이퍼 파라미터 최적화

- 최적화란 함수 $f: X \rightarrow R$ 에 대해 $f(x)$ 를 최소로 하는 해를 찾는 방법을 의미한다. 그 중 가장 좋은 성능 측정값을 갖는 하이퍼파라미터를 선정하는 방식을 하이퍼파라미터 최적화라고 한다(Bergstra et al., 2012).
- 일반적으로 최적의 성능 측정값을 갖는 하이퍼파라미터는 아래 식으로 표현할 수 있다.

$$\lambda^* \equiv argmin\ \Psi(\lambda) \tag{21}$$

λ^*: 최적해
λ: 하이퍼파라미터 셋
Ψ: 하이퍼파라미터 반응함수(목적함수)

- 하지만, $\lambda \in \Lambda$에 대하여 하이퍼파라미터가 존재하는 모든 실수 범위에 대해 Ψ를 평가하는 것은 불가능하다.
- 따라서, 현실적으로 하이퍼파라미터 반응함수를 평가가능한 정도의 적당한 자연수 S를 선정하여 $\lambda \in \{\lambda^{(1)}, ..., \lambda^{(S)}\}$에 대하여 $\lambda^* \approx \hat{\lambda}$인 최적 하이퍼파라미터 추정값 $\hat{\lambda}$를 찾게된다.
  - 이때, Ψ는 일반화된 성능측정값을 나타내기 위해 교차검증(Cross - validation)을 사용한다(Bergstra et al., 2012).
- 하이퍼파라미터를 최적화하는 알고리즘은 다양하게 존재한다. 가장 많이 사용되는 방식은 수동 선택법(Manual search), 격자 탐색법(Grid search), 랜덤 탐색법(Random search)이 있다.
  - 수동 선택법은 가장 기본적인 방식으로 실무자의 경험과 사전 지식에 따라 하이퍼파라미터를 선택하여 평가하고 $\hat{\lambda}$를 찾는 방식으로 불안정하고, 하이퍼파라미터의 차원이 높은 모델의 경우에는 적용이 거의 불가능하다는 단점이 있다.
  - 격자 탐색법은 수동 선택법의 불안정성을 해결하기 위해 각 하이퍼파라미터별로 구간을 임의의 격자로 나누어, 해당 파라미터의 구간을 이산형 집합으로 정의한 뒤, 각 하이퍼파라미터의 곱집합으로 표현되는 경우의 수에 대해 반응함수를 평가하여 $\hat{\lambda}$를 찾는다.

- 이 방식은 하이퍼파라미터의 수가 증가하면 평가에 필요한 비용이 기하급수적으로 증가한다는 문제점과 지정된 경우의 수에 포함되지 않는 격자사이의 값은 반응함수 평가에 포함되지 않는다는 한계점이 있다(Bellman, 1961).
- 랜덤 탐색법은 임의의 수 S만큼 하이퍼파라미터 공간상에서 무작위로 지점을 선정하여 반응함수를 평가하므로 격자 탐색법의 한계점을 보완하면서도 효과적으로 $\hat{\lambda}$를 찾는다(Bergstra et al., 2012).
- 그러나, 무작위성에 의존하여 하이퍼파라미터를 탐색하는 방식은 하이퍼파라미터 수가 많은 경우 적은 평가횟수로는 신뢰도가 낮아 많은 평가횟수를 선정해야한다는 문제점과 전역 최적해를 보장하는 논리적 신뢰도가 부족하다는 한계점이 있다.

- 이에 본 연구에서는 베이지안 최적화를 활용하여 딥러닝 모델의 하이퍼파라미터 최적화를 수행하고, 모델의 성능을 확보하고자 한다.
  - 베이지안 최적화는 최적해 탐색과정에서 이전 하이퍼파라미터 탐색결과를 이용하여, 확률모델 기반으로 다음 탐색지점을 선정하기 때문에 가장 효율적으로 하이퍼파라미터를 탐색하는 알고리즘으로 알려져 있어, 목적함수 평가에 비용이 많이 드는 블랙박스 모델의 하이퍼파라미터 탐색에 적용하기 적합하다고 판단된다(Bergstra et al., 2011; Mockus, 2012).

✓ **베이지안 최적화(Bayesian optimization)**

- 베이지안 최적화는 수치최적화(Numerical optimization)의 한 일종으로 큰 틀에서는 시행착오를 통해 최적화 조건을 만족하는 최적해를 스스로 탐색하는 수치최적화와 과정이 동일하다.
- 베이지안 최적화는 모델기반으로 시행착오 과정에서 다음 탐색할 해를 결정하는 부분에서 차별점이 있으며, 이와 같은 특성 때문에 순차 모델 전역최적화(Sequential Model-Based Global Optimization, SMBO)라고도 한다.
- 베이지안 최적화는 선택함수(selection function 또는 Acquisition function)을 통해 순차적인 하이퍼파라미터 탐색과정에서 이전 탐색결과들을 토대로 다음 탐색지점을 선정한다.
- 선택함수는 다양한 종류가 있으며, 그 중 Expected Improvement(EI)는 함수가 의미하는 바가 직관적으로 이해하기 쉽고 다양한 상황에 적용하여도 그 성능이 안정적이라는 장점을 갖고 있으며, 아래와 같이 표현된다.

$$EI_{y^*}(x) = \int_R \max(y^* - y, 0) p_m(y|x) dy. \tag{22}$$

y^*: 기존 탐색값 중 임의의 임계값(보통 하위 15%)
y: 하이퍼파라미터 x에 대한 확률분포 p_m상의 성능측정값
x: 하이퍼파라미터 탐색값
p_m: 하이퍼파라미터 x에 대한 성능측정값 y의 확률분포(surrogate model)

- 즉, 성능 측정값의 기댓값이 크거나(p_m) 하이퍼파라미터가 x일 때 성능측정값의 불확실성이 큰 경우 EI 가 큰 값을 갖게 되므로 베이지안 최적화는 순차적 탐색과정에서 성능측정값이 컸던 지점 또는 확률분포상 불확실정이 큰 지점에 큰 비중을 두고 순차적 탐색과정을 진행한다는 것을 의미한다.
- 다수의 하이퍼파라미터를 지닌 딥러닝 모델의 특성상 하이퍼파라미터 값에 따라 도출되는 성능측정값의 함수 $f : X \rightarrow Y$는 구조를 알기 어렵고 알아내려는 과정자체가 비현실적이기 때문에 대안으로 확률모델인 Surrogate model(p_m)을 활용한다.
- 본 연구에서는 surrogate model로 TPE(Tree Parzen Estimator)를 사용하며, TPE에서는 $p_m(x|y)$을 다음과 같이 표현할 수 있고, 이때 y^*는 이전 탐색결과 중 임의의 값(주로 하위 15%)을 의미한다.
  - 대표적인 surrogate 모델로 gaussian process(GP)와 TPE가 있으며, 다양한 선행 연구 사례에서 TPE가 더 우수한 성능을 나타낸 바 있어 TPE를 사용하고자 한다.

$$p_m(x|y) = \begin{cases} l(x) & if \ \ y < y^* \\ g(x) & if \ \ y \geq y^* \end{cases} \tag{23}$$

$l(x)$: y^*이하일 때 TPE의 surrogate model
$g(x)$: y^*이상일 때 TPE의 surrogate model

- 따라서, 식(19)에 따라 식(18)은 아래와 같이 표현할 수 있다.

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(x|y) \frac{p(y)}{p(x)} dy \tag{24}$$

- 그리고, r과 $p(x)$는 아래와 같이 정의할 수 있다.

$$r = p(y < y^*) \tag{25}$$

$$p(x) = \int_{-\infty}^{+\infty} p(x|y)p(y)dy = rl(x) + (1-r)g(x) \tag{26}$$

- 정리하면, 아래와 같이 표현할 수 있다.

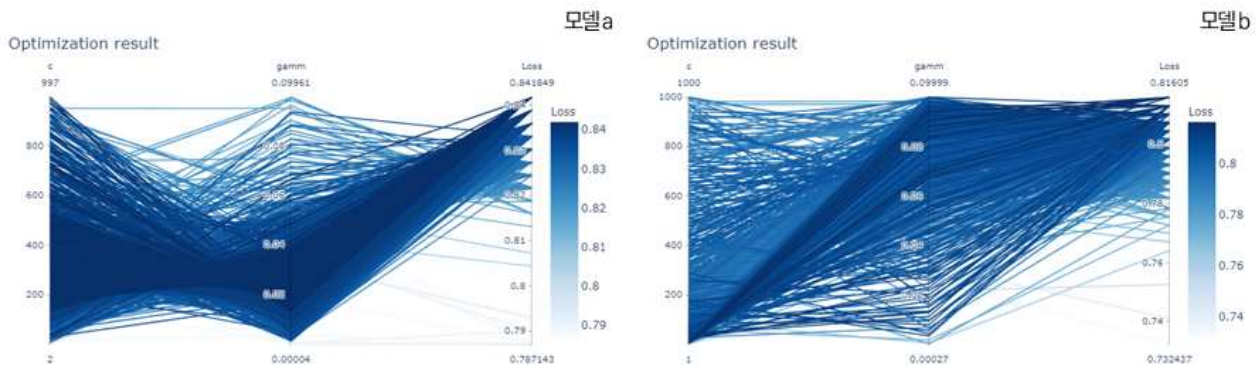
$$\int_{-\infty}^{y^*} (y^* - y)p(x|y)p(y)dy = l(x)\int_{-\infty}^{y^*} (y^* - y)p(y)dy \tag{27}$$

$$l(x)\int_{-\infty}^{y^*} (y^* - y)p(y)dy = ry^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy$$

- 최종적으로 식(23)은 아래와 같이 정리될 수 있다.

$$EI_{y^*}(x) = \frac{ry^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy}{rl(x) + (1-r)g(x)} \propto (r + \frac{g(x)}{l(x)}(1-r))^{-1} \tag{28}$$

- 즉, EI는 $l(x)$에 비례하고 $g(x)$에 반비례하며, 이는 TPE가 최적화 문제에서 특정 임계값 r보다 더 좋은 성능을 보인 탐색결과에 더 가중치를 두고 최적화를 진행하는 것을 의미한다 (Bergstra et al., 2011).

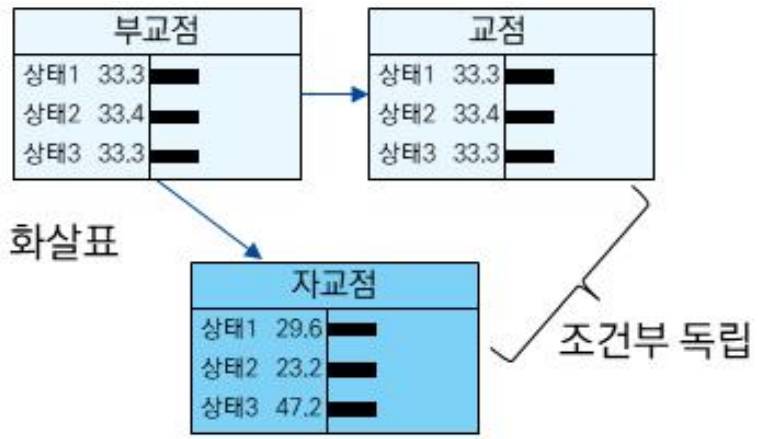


<그림 2-30> TPE를 활용한 하이퍼파라미터 최적화 예시

2). 조건부확률 모델 구축

✓ 베이지안 네트워크(Bayesian networks)

- 베이지안 네트워크는 해석적인 측면에서 우수할 뿐만 아니라, 변수 간의 상호 관계 혹은 예측에 대한 불확실성에 대해 조건적인 의존 관계를 확률적으로 표현하므로 환경 정책 등 의사결정을 지원에 있어 매우 적절한 방법이다.
- 베이지안 네트워크는 도식적 모델로 변수 사이의 상호 인과관계를 기반으로 해석 가능한 모델 구조를 제공한다.
  - 베이지안 네트워크는 변수의 상태를 나타내는 교점과 변수들 간 연결을 나타내는 화살표로 구성되는 도식적 모델이다.
  - 두 교점이 화살표로 이어져 있다면, 부교점과 자교점 간의 확률적 관계가 조건부 확률표(Conditional Probability Table, CPT)로 표현되는 의존성이 있음을 나타낸다.
  - 반면, 두 교점이 화살표로 이어져 있지 않다면, 두 변수는 조건부 독립성을 나타낸다.



<그림 2-31> 베이지안 네트워크의 구조

- 따라서, 특정 자교점 X_i가 부교점의 집합 $pa(X_i)$에 의해 확률적으로 표현된다고 할 때, 자교점의 조건부 확률 분포는 $P(X_i|pa(X_i))$로 표현이 된다.
- 이때, 전체 랜덤 변수에 대한 확률 분포는 결합 확률 분포(Joint probability distribution)라 하며, 베이지안 네트워크상 조건부 확률 분포를 고려한 결합 확률 분포 $P_X(X)$는 아래와 같이 표현할 수 있다.

$$P_X(X) = \prod_{i=1}^{p} P_{X_i}(X_i|pa(X_i)) \tag{29}$$

- 베이지안 네트워크의 구축 과정은 크게 데이터 전처리와 학습으로 구분할 수 있으며, 전처리 과정은 변수 이산화(discretization), 변수 선택(variable selection)을 포함하고, 학습은 구조 학습(structure learning) 및 파라미터 학습(parameter learning)을 포함한다.

✓ **자료 전처리**

- 변수 이산화과정은 연속적인 변수를 N개의 이산화된 범주로 나누는 것을 의미한다.
  - 이산화를 통해 생성된 범주의 개수가 너무 적거나 많은 경우, 분포의 의미를 손실하거나 모델의 복잡성이 급수 적으로 증가해 모델의 성능을 저하 시킬 수 있으으로 현실적인 범위가 필요하다.
  - 변수 이산화를 위한 방법 대표적인 방식은 범위법, 빈도법, 상호 정보법의 크게 세 가지 방식으로 구분되며, 전문가의 의견 수렴을 통한 방식도 사용된다.
  - 범위법(Interval-based method)은 연속변수를 동일한 범위로 이산화하는 방법으로 적용이 간단하나, 연속변수 분포의 왜도가 클 경우, 효과적이지 못하며, 관련된 변수 사이의 관계성이 고려되지 않는 한계가 있다.
  - 빈도법(Quantile-based method)은 연속변수를 빈도를 기반으로 이산화하는 방법으로 범위법 보다 변수의 분포 모양을 고려한 방법이나, 관련된 변수 사이의 관계성은 고려하지 못하는 한계가 있다.
  - 상호 정보법(Mutual information-based method)은 연속 변수 사이의 상호 정보를 기반으로 이산화하는 방법으로 Hartmink 방법이라고도 부르며(Hartmink, 2001), 범위법이나 빈도법보다 변수의 분포를 잘 표현하는 방법이지만, 복잡한 관계를 가지고 있는 다수의 변수에 대해 적용하기에 한계가 있다.
- 변수 선택 과정은 다양한 변수 중 모델의 종속변수와 상당한 관련이 있는 것으로 판단되는 주요변수들을 선택하는 과정이다.
  - 베이지안 네트워크에서 너무 많은 입력변수는 모델의 복잡도를 증가시키며, 과적합과 성능의 감소로 이어질 수 있으며, 너무 적은 독립변수는 입력변수를 설명하기에 충분한 정보가 되지 않을 수 있다.
  - 변수 선택 과정의 대표적인 방법은 필터 방법, 래퍼 방법, 임베디드 방법의 크게 세 가지로 구분할 수 있다.

- 필터 방법(Filter method)은 종속변수와 입력변수 간의 상관계수, 상호정보 등과 같은 단순 통계량을 바탕으로 변수를 선택하는 방법으로 단순하고 간단히 구현할 수 있는 방법이나, 비슷한 관계성을 나타내는 변수들을 구분하지 못해 다중공선성을 해결하지 못한다.
- 래퍼 방법(Wrapper method)은 방법으로 다양한 입력변수의 조합을 통해 모델의 성능을 평가한 뒤, 가장 좋은 모델 성능을 보이는 대입 변수의 조합을 선택한다.
- 모든 대입 변수 경우의 수 조합에 대해 수행하거나 전진선택법(Forward selection), 후진소거법(Backward elimination), 유전알고리즘(Genetic algorithm) 등의 방법이 사용되며, 다중공선성에 대한 문제는 해결할 수 있으나, 시간과 비용적인 측면에서 구현하기 상대적으로 어렵다.
- 임베디드 방법(Embedded method)은 다른 모델을 사용하여 특정 목적함수를 최적화하거나 모델 퍼포먼스의 증가를 통해 중요 변수를 추출하는 방식으로 주로 L1 회귀분석(Lasso regression)이나 의사결정 나무를 활용한다.
- 다중공선성과 시간 및 비용적인 측면을 동시에 해결하여 탁월한 방법이나, 실제 분석에 사용할 모델 방법과는 다른 모델을 통해 중요한 변수를 선택하기 때문에 해석적인 측면에서 주의가 필요한 방법이다.

✓ **베이지안 네트워크 학습**

- 베이지안 네트워크의 학습은 구조 학습과 파라미터 학습으로 이루어지며, 구조 학습은 입력변수 간 인과관계 및 그 방향성을 설정하는 과정이며, 파라미터 학습은 구조 학습을 통해 결정된 조건부 확률표를 최적화하는 과정이다.
- 즉, 베이지안 네트워크의 학습은 다음과 같은 구조로 표현할 수 있다.

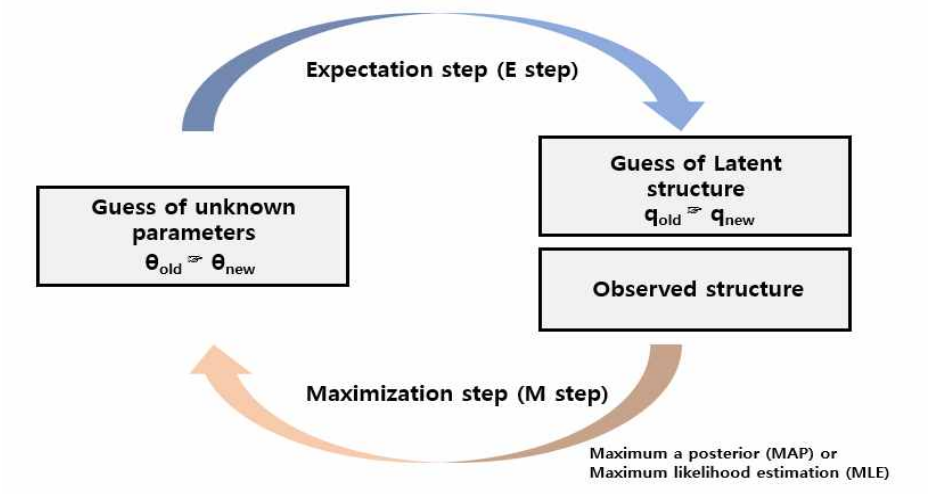
$$\underbrace{P(B|D) = P(G,\Theta|D)}_{learning\ \ BNs} = \underbrace{P(G|D)}_{Structure\,learning} \bullet \underbrace{P(\Theta|G,D)}_{Parameter\,learning} \tag{30}$$

D : 데이터

$B = (G, X)$: 베이지안 네트워크

Θ: 입력변수의 집합 X의 분포를 설명하는 모수의 집합

- 구조 학습은 선행 연구사례 및 전문가 자문 등을 활용하여 구조를 구성하는 지식 기반 구조 학습과 데이터에 대해 알고리즘을 적용하여 구조를 학습하는 데이터 기반 구조 학습 방식으로 구분할 수 있다.
  - 지식 기반 구조 학습은 지속적인 전문가들의 자문을 바탕으로 입력변수 사이의 인과관계를 결정하는 방법으로, 현재까지 인정된 지식 체계를 반영할 수 있으며, 데이터가 부족한 상황이거나 여러 오차에 의한 불확실성이 커 데이터 상 노이즈가 심할 경우 효과적으로 활용할 수 있다.
  - 데이터 기반 구조 학습은 다양한 알고리즘을 통해 자동적으로 랜덤 변수 사이의 인과 관계를 결정하는 방법이며, 새로운 인과 관계의 가능성에 대해 발견할 수 있어 확장적인 장점이 있다.
  - 그러나, 사용되는 알고리즘마다 결과로 나타나는 구조에 많은 차이가 있고, 입력변수의 분포에 대해 통계적 유의성을 판단 및 검증하여야 올바른 인과 구조의 생성이 가능하다.
  - 구조 학습 알고리즘은 크게 인과 관계의 통계적 가설 검정을 통해 구조를 형성하는 Constraint-based 알고리즘(PC, Grow-Shrink(GS), Incremental Association(IAMB) 등), 휴리스틱 한 방식으로 전체 모델 구조의 목적함수 수치를 최적화하는 Score-based 알고리즘(Greedy search, Genetic algorithm, Simulated annealing 등), 두 가지 방식을 동시에 사용하는 Hybrid 알고리즘(Sparse candidate algorithm(SC), Max-Min Hill-Climbing algorithm(MMHC)) 세 가지로 분류할 수 있다.
- 파라미터 학습은 데이터를 기반으로 랜덤 변수의 조건부 확률표를 최적화하는 과정으로 주로 Expectation-Maximization 알고리즘(EM 알고리즘)이나 경사 하강법(Gradient decent)으로 수행된다. 일반적으로 EM 알고리즘이 더 모델의 견고함이(robustness)이 높다고 알려져 있다.
  - EM 알고리즘은 관측되지 않은 잠재 변수에 의존하는 확률 모델에서 관측 변수의 분포에 대해 최대 가능도(Maximum likelihood)나 최대 사후 확률(Maximum a posteriori)을 갖는 매개변수를 반복적으로 찾는 알고리즘이다.

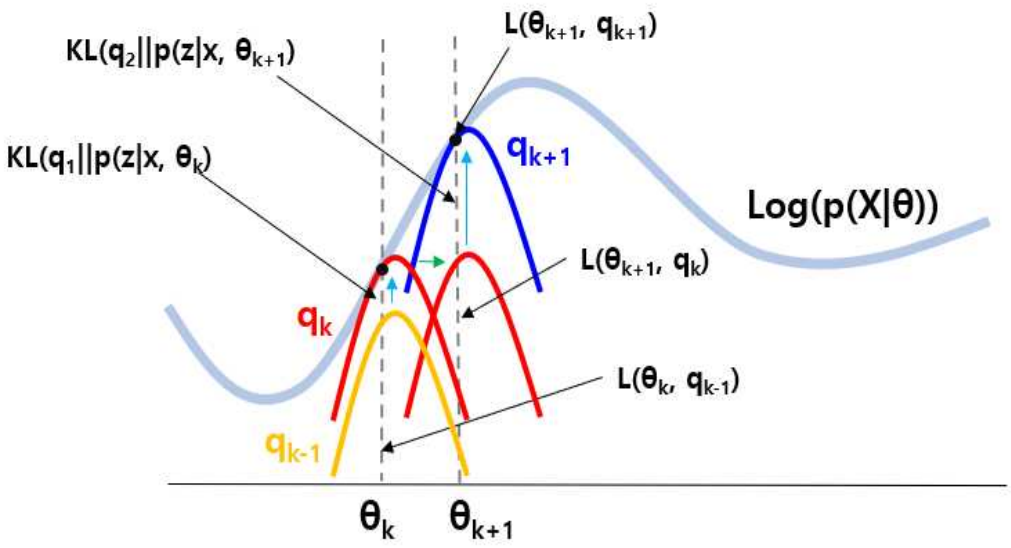


<그림 2-32> 간략화한 EM 알고리즘의 적용 과정

- EM 알고리즘은 파라미터로부터 잠재적 구조를 추론하는 Expectation step(E-step)과 관측된 구조를 통해 파라미터의 가능도나 사후확률을 최대화하여 갱신하는 Maximization step(M-step) 구조로 구성된다. EM 알고리즘의 최종 목적은 가능도를 최대화하는 잠재 변수에 대한 확률 분포와 모수를 구하는 것이 목표이다.

$$\max_{\theta} \log(p(X|\theta)) = \sum_{i=1}^{N} \log(p(x_i|\theta)) \tag{31}$$

θ : 모수, X : 무작위 변수

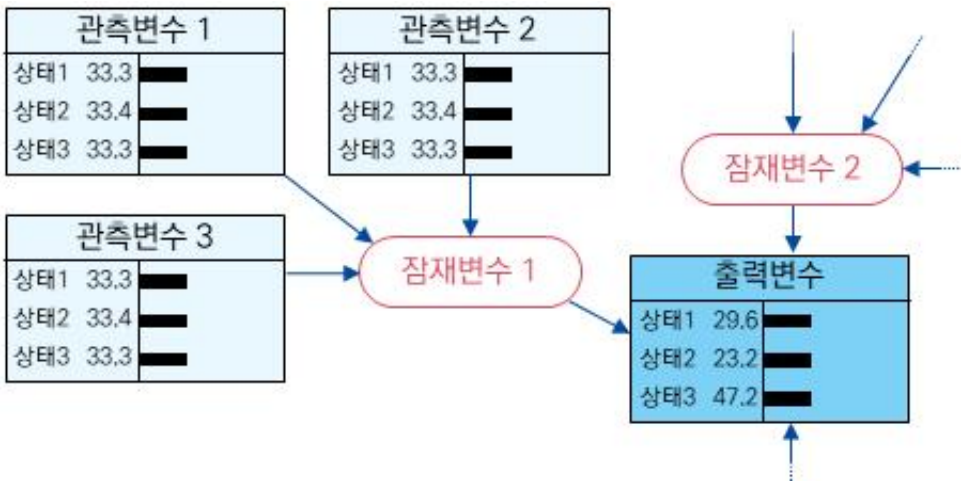


<그림 2-33> EM 알고리즘의 개념도

- EM 알고리즘의 수렴할 때까지 E와 M 단계를 반복적으로 수행되며 이러한 반복은 새로운 모델의 변수가 기존 모델보다 부적합할 때 멈추게 된다.
- 이때 두 가지 멈춤 조건(stopping criterion)을 사용할 수 있으며, 1) 가능도 함수의 차가 수렴하여 거의 차이가 없거나 2) 모수의 차이가 유의하지 않을 경우 EM 알고리즘은 종료된다.

✓ **계층적 베이지안 네트워크(Hierarchical Bayesian network)**

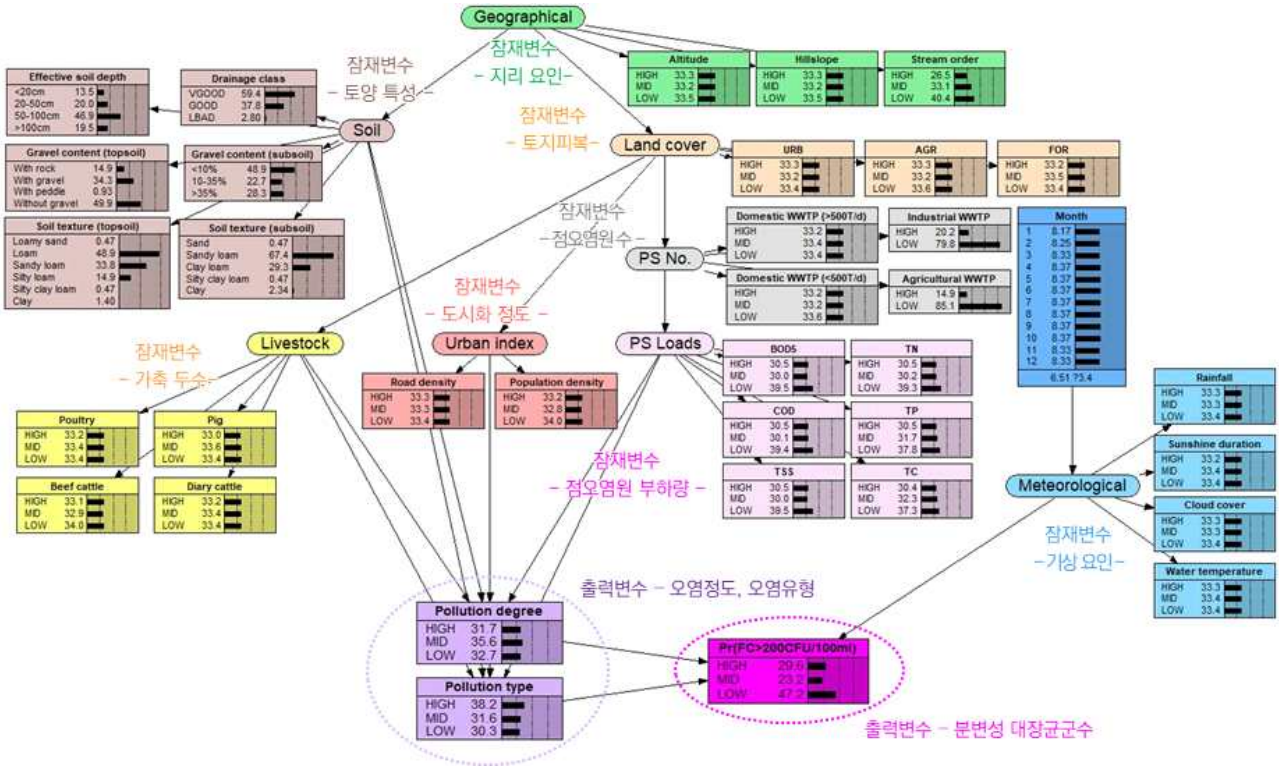
- 일반적인 베이지안 네트워크에서 입력변수 수의 증가는 중요 입력변수들의 변화에 대한 종속변수의 민감도를 감소시키며, 결과적으로 과적합(Overfitting)을 일으키고 모델 성능을 감소시킨다.
- 입력변수의 증가에 따른 문제는 변수 선택 등을 해결할 수 있으나, 이는 부득이하게 정보의 손실을 야기한다.
- 변수 선택에 대안적인 방법으로 베이지안 네트워크 내부에서 잠재적인 변수(Latent variable)를 활용하여 모델을 구조화하는 방법이 있으며, 이와 같이 잠재 변수를 활용해 구조화한 베이지안 네트워크를 계층적 베이지안 네트워크라 한다.



<그림 2-34> 계층적 베이지안 네트워크의 구조

- 계층적 베이지안 네트워크의 잠재변수는 유사한 관측변수(observed variable) 간의 정보를 압축, 공통적이며 대표적인 정보를 추출함으로써 특정 변수 그룹의 일반화된 정보를 제공한다.

- 따라서, 계층적 베이지안 네트워크는 수 많은 입력변수를 필요로 하는 복잡한 환경 현상을 표현함에 있어 변수 선택 및 제거 등으로 인한 정보손실을 방지하면서도 복잡하고 다중적인 계층 구조에 대해 효과적으로 표현할 수 있으며, 관측변수 및 잠재변수 간의 상호작용 효과 등을 예측 가능하고 과적합이나 교란 관계(spurious relationship)에 따른 잠재적인 함정을 회피할 수 있다.



<그림 2-35> 분변성 오염에 대한 계층적 베이지안 네트워크 구축 사례

2-2-2-4. 구축된 평가모델에 대한 예측성능 평가

- 최적화 과정 이후 구축이 완료된 평가모델들은 분류모델에 대해 다양한 평가지표를 활용하여 예측성능에 대한 검증을 수행한다.
- 분류모델의 예측결과는 주로 혼동행렬(confusion matrix)으로 표현되며, 예측결과에 대한 성능지표는 정확도(accuracy), 정밀도(precision) 및 재현율(recall), F1 score, AUC(Area under reciever operating characteristic curve) 등이 있다 <표 2-8>.

<표 2-8> 혼동행렬의 형태

Confusion matrix		Observed class	
		True	False
Predicted class	Positive	True positive (TP)	False positive (FP, 일종 오류)
	Negative	False negatives (FN, 이종 오류)	True negatives (TN)

- 정확도는 전체 샘플 수에 대한 예측을 통해 얻어진 분류 범주가 실제의 분류 범주와 같은 샘플 수의 비율을 의미하는 가장 직관적이고 간편한 성능 평가방법이다.
- 다만, 분류 문제에서 불균형 데이터(Imbalanced data)의 경우에 희소한 경우에 대한 모델 성능을 쉽게 평가하지 못한다.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{32}$$

- 정밀도는 모델이 positive라고 분류한 것 중에서 실제 positive인 샘플 수의 비율을 의미하며 재현율은 전체 positive 샘플 수에 대한 모델에서 positive이라고 예측한 샘플 수의 비율을 의미한다.
- 정밀도와 재현율은 공통적으로 TP(true positive)가 분자이며, 분모에서 TP를 공통으로 활용함에 더해 제 1종 오류와 제 2종 오류에 해당하는 FN(false negative)과 FP(false positive)를 가지고 있으므로 정밀도와 재현율은 서로 trade-off 관계에 있는 것을 확인할 수 있다.

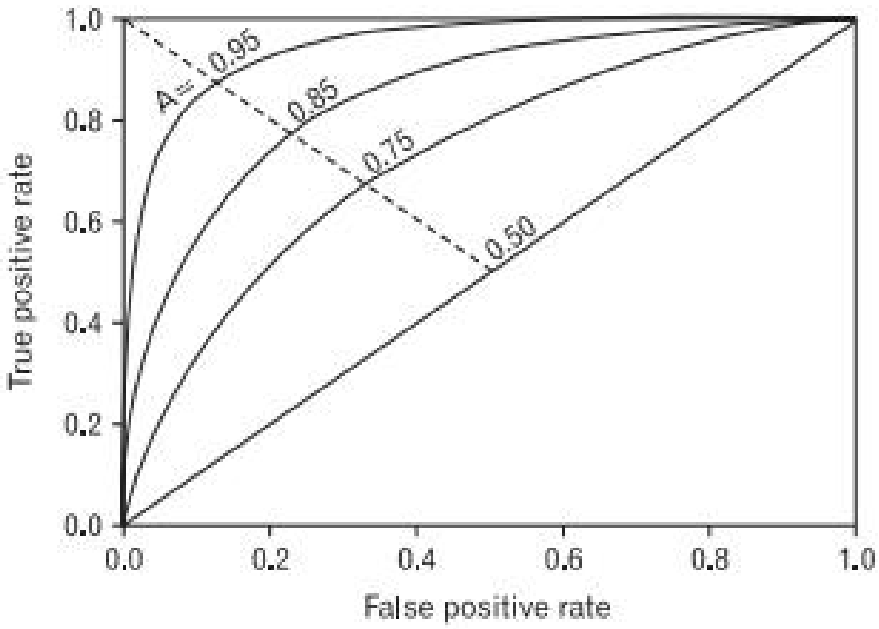
$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \tag{33}$$

$$Recall = \frac{TP}{TP + FN} \tag{34}$$

- F1 score는 정밀도와 재현율의 조화 평균으로 불균형 데이터 존재할 때 제 1종 오류 및 제 2종 오류를 모델 성능 지표에 동시에 반영할 수 있는 장점이 있다.

$$F-measure = 2 \times \frac{1}{\frac{1}{\mathrm{Pr}ecision} + \frac{1}{Recall}} = 2 \times \frac{\mathrm{Pr}ecision \times Recall}{\mathrm{Pr}ecision + Recall} \tag{35}$$

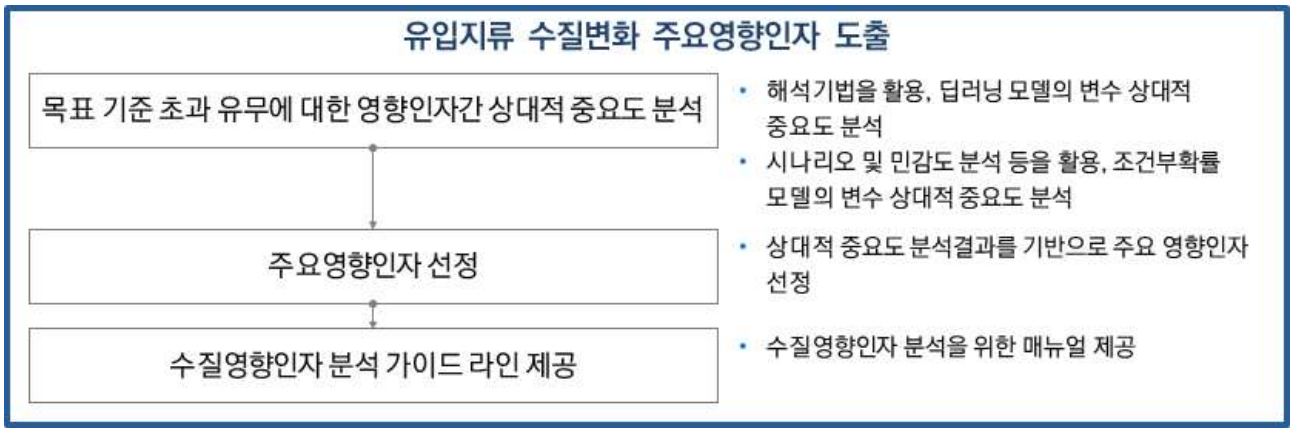
- AUC는 모든 임계값에서 분류모델의 성능을 민감도(true positive rate)와 특이도(true negative rate)를 한꺼번에 반영하여 나타내는 지표로 x축을 false positive rate(1-특이도), y축을 true positive rate으로 하는 ROC(Receiver operating characteristics) 커브의 아래 면적을 의미하며, 우수한 모델일수록 AUC의 값이 1에 가까우며 최솟값은 0.5로 모델의 분류 능력이 없음을 의미한다 <그림 2-36>.
- 이때, 민감도(sensitivity)는 TP/(TP+FN)로 산정이 가능하고, 특이도(specificity)는 TN/(FP+TN)와 같다.



<그림 2-36> ROC 커브 예시

2-2-3 유입지류 수질변화 주요영향인자 도출

- 유입지류 수질변화 주요영향인자 도출 단계로 상대적 중요도 분석결과 종합, 주요영향인자 선정을 통해 수질영향인자 분석 가이드 라인 제공으로 구성한다 <그림 2-37>.
  - 목표 기준미달성에 대한 상대적 중요도 분석 단계에서는 딥러닝 모델의 경우 모델 해석기법 등을 활용하며, 조건부확률 모델의 경우 시나리오 분석 및 민감도 분석 등을 활용하여 수행한다.
  - 상대적 중요도 분석결과 종합 단계에서는 딥러닝 모델 및 조건부확률 모델에 대한 변수 상대적 중요도 분석결과를 다각도로 분석, 결과를 종합한다.
  - 이 과정에서 강우 등 영향인자가 모델의 예측 결과(목표기준 달성/미달성) 상대적 중요도, 기여도 등을 심층적으로 분석하며, 영향인자의 변화에 따른 예측 결과의 차이 등을 함께 분석한다.
  - 주요 영향인자 선정 단계에서는 앞선 단계를 통해 통합된 분석결과를 기반으로 주요 영향인자를 선정한다.
  - 수질영향인자 분석 가이드 라인 제공 단계에서는 과업을 통해 구축된 모델 및 그를 활용한 주요영향인자 도출을 위한 매뉴얼을 작성한다.



<그림 2-37> 유입지류 수질변화 주요영향인자 도출 흐름도(안)

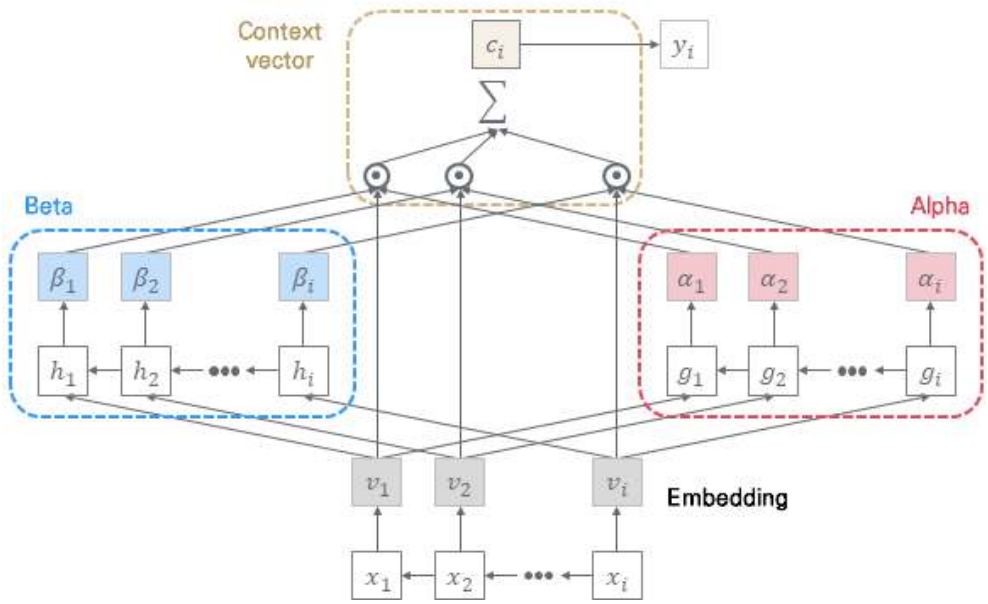
2-2-3-1. 모델 해석기법을 활용한 영향인자 간 상대적 중요도 분석

✓ **어텐션 매커니즘 (Attention mechanism)**

- 어텐션 메커니즘 (Attention mechanism)은 모델학습 시 관심 영역의 가중치를 높여 학습하며, 디코더(decoder)에서 출력 결과를 예측하는 시점마다 인코더(ecoder)의 은닉층(hidden layer)을 다시 한번 참고하는 방법으로 어떤 인코더의 은닉층을 얼마나 참고할지 결정하는 어텐션 스코어(attention score)를 기반으로 어텐션 가중치(attention weight)를 산정한다.
- 어텐션 메커니즘에서 사용되는 어텐션 가중치는 특정 값을 얼마나 가중 시킬 것인지를 의미하는 가중치와 유사하나 어텐션 가중치가 전체 또는 특정 영역의 입력값을 반영하여 그중 어떤 부분에 집중하는지를 나타낸다는 점에서 차이가 있다.
- 어텐션 매커니즘을 사용하면 정보손실을 최소화할 수 있고 순환신경망(RNN) 또는 장단기 인공신경망(LSTM)에서 발생하는 기울기 소실 문제를 해결할 수 있다.

✓ **시간역순 어텐션 매커니즘 (REverse Time AttentIoN mechanism)**

- 기존의 인공신경망은 결과에 대한 해석이 어려운 블랙박스 모델이라는 한계가 있다.
- RETAIN은 해석 가능한 인공신경망 중 하나로 어텐션 매커니즘을 기반으로 시계열 데이터에서 시점 정보와 변수 정보 각각에 대해 어텐션 가중치를 산정함에 따라 데이터 정보 손실을 방지할 수 있다 <그림 2-38>.



<그림 2-38> RETAIN 모델의 구조

- RETAIN 과정은 시점별, 변수별 해석을 고려하는 과정을 포함하여 4단계로 나뉘며 시점별 변수의 기여도 확인이 가능하다.
- RETAIN의 입력 데이터는 임베딩(embedding)과정을 통해 수치화할 수 있으며, 선형 임베딩을 사용할 경우 아래와 같은 식으로 표현할 수 있다.

$$v_i = W_{emb} x_i \tag{36}$$

W_{emb} : 학습하는 임베딩 지표
v_i : 입력벡터의 임베딩
x_i : 입력벡터

- 입력 데이터를 기반으로 두 개의 순환신경망을 통해 α, β값을 산정한다. 시계열 데이터에서 시점 해석을 보존하기 위해 알파 순환신경망(RNN_α)을 통해 구한 어텐션 가중치로 소프트맥스 함수를 사용하여 어텐션 분포 α값을 계산하며 각 시점을 임베딩한 값을 $v_1,...,v_i$이라한다. 식은 아래와 같이 표현할 수 있다.

$$g_i, g_{i-1}, ..., g_1 = RNN_\alpha(v_i, v_{i-1}, ..., v_1) \tag{37}$$

g_i : 알파 순환신경망 은닉계층
v_i : 입력벡터의 임베딩

$$e_j = w_\alpha^\top g_j + b_\alpha \quad \text{for } j = 1, ..., i \tag{38}$$

e_j : 디코더, 인코더 은닉 상태의 어텐션 스코어 모음값
w_α, b_α : 학습 파라미터
g_i : 알파 순환신경망 은닉계층

$$\alpha_1, \alpha_2, ..., \alpha_i = Softmax(e_1, e_2, ..., e_i) \tag{39}$$

α_i : 어텐션 분포
e_i : 디코더, 인코더 은닉 상태의 어텐션 스코어 모음값

- 이때, 변수 해석 보존을 위해 베타 순환신경망(RNN_β)을 통한 어텐션 가중치 산정 및 어텐션 분포 β값을 계산하는 과정이 함께 이루어지는데 β값 산정 부분은 하이퍼볼릭 탄젠트 함수를 사용하여 아래와 같이 표현할 수 있다.

$$h_i, h_{i-1}, ..., h_1 = RNN_\beta(v_i, v_{i-1}, ..., v_1) \tag{39}$$

h_i : 베타 순환신경망 은닉계층
v_i : 입력벡터의 임베딩

$$\beta_j = \tanh(W_\beta h_j + b_\beta) \text{ for } j = 1, ..., i, \tag{40}$$

β_j : 어텐션 분포
W_β b_β : 학습 파라미터
h_j : 베타 순환신경망 은닉계층

- RETAIN은 순환신경망에 입력 데이터를 시간 역순으로 적용하여 어텐션 가중치를 생성하므로 α, β 계산 과정에서 각 시점의 임베딩 값을 $v_{i-1}, ..., v_1$로 적용한다.
- 시간 역순으로 데이터를 입력하게 되면, 각 시간간격(time step)에서 값을 예측할 때 순환신경망 층의 결과값 변화를 유발하고 예측값 생성을 위해 새로운 세트의 α, β를 생성하므로 어텐션 벡터가 각 시간간격마다 다르게 출력된다. 시간 역순이 아닌 정순으로 넣을 경우 α_1과 β_1는 항상 동일한 값을 갖는다.
- 생성된 시점별, 변수별 어텐션 가중치를 통해 어텐션 함수의 출력값인 컨텍스트 벡터(context vector)를 생성하는 부분은 아래와 같이 표현할 수 있다.

$$c_i = \sum_{j=1}^{i} \alpha_j \beta_j \odot v_j \tag{41}$$

c_i : 컨텍스트 벡터
α_j : 시점별 어텐션 분포
β_j : 변수별 어텐션 분포
$\odot$: 요소별 곱
v_j : 입력벡터의 임베딩

- 이를 토대로 모델의 예측값은 소프트맥스 함수와 손실함수로 교차 엔트로피를 사용하여 아래와 같이 표현할 수 있다.

$$\hat{y_i} = Softmax(Wc_i + b) \tag{42}$$

$\hat{y_i}$: y 예측값

W_{c_i} b : 학습 파라미터

- 이처럼 RETAIN은 시점(α_j)과 변수(β_j) 모두를 고려한 어텐션을 가지고 있으며 특정 시점에서 입력 데이터의 최종 기여도를 계산할 때 두 가지를 모두 고려한다.
- 이를 수식으로 표현하면 아래와 같으며, j 시점에서 입력 데이터의 k번째 변수의 기여도는 기여계수와 입력값 $x_{j,k}$의 곱으로 표현된다.

$$w(y, x_{j,k}) = \alpha_j W(\beta_j \odot W_{emb}[:,k])\, x_{j,k} \tag{43}$$

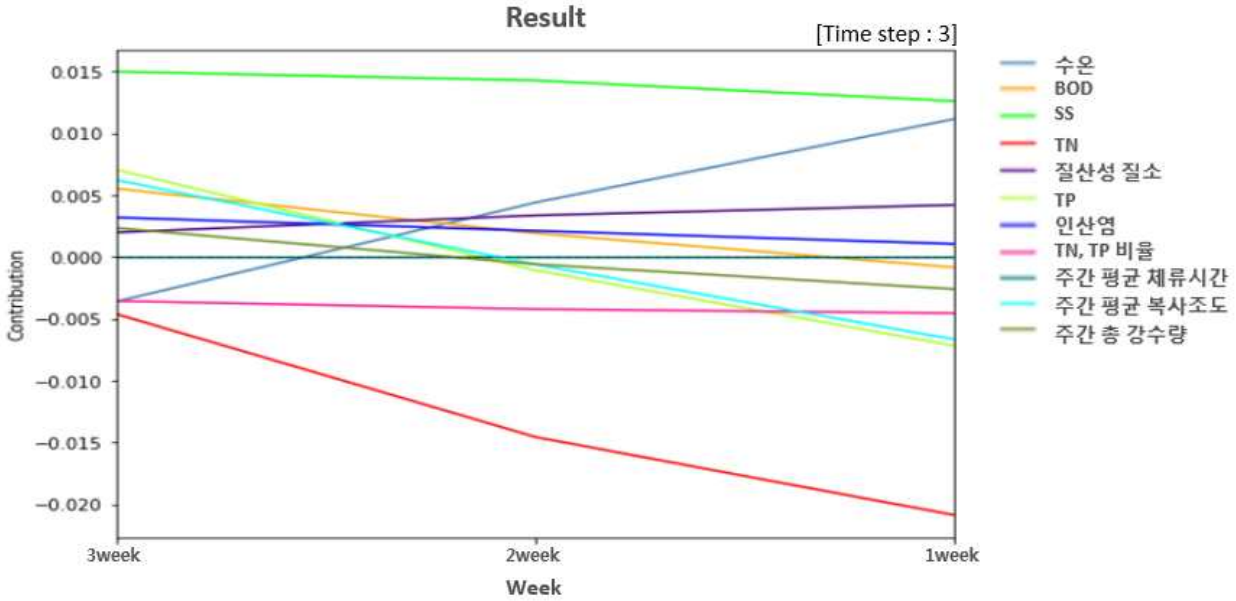
w : 기여도

α_j : 시점별 어텐션 분포

β_j : 변수별 어텐션 분포

W : 학습 파라미터

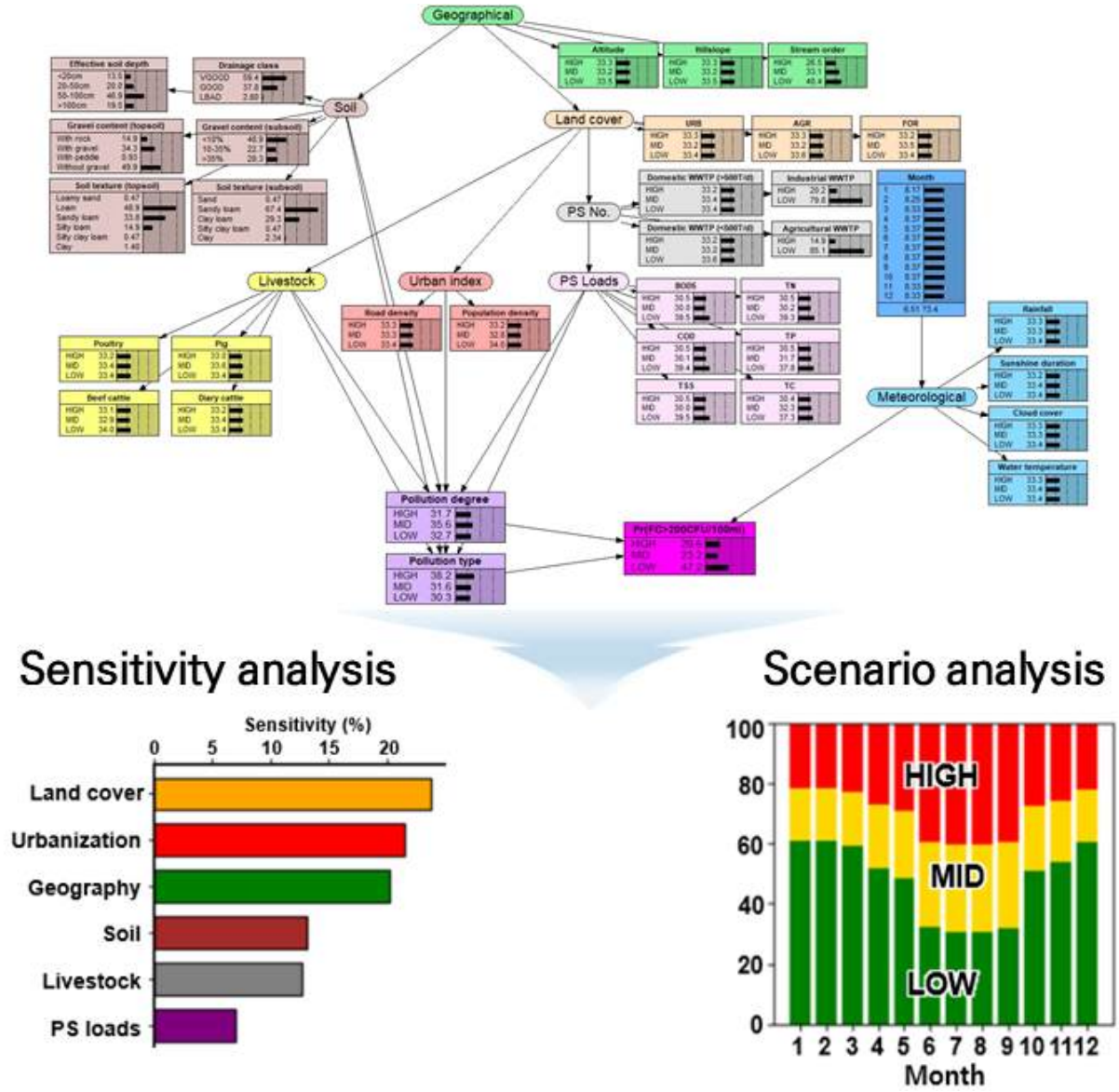
$x_{j,k}$: 입력값



<그림 2-39> RETAIN 모델을 활용한 유해남조류 세포 수 예측에 대한 환경인자 기여도 분석 예시

2-2-3-2. 모델 특성을 활용한 영향인자 간 상대적 중요도 분석

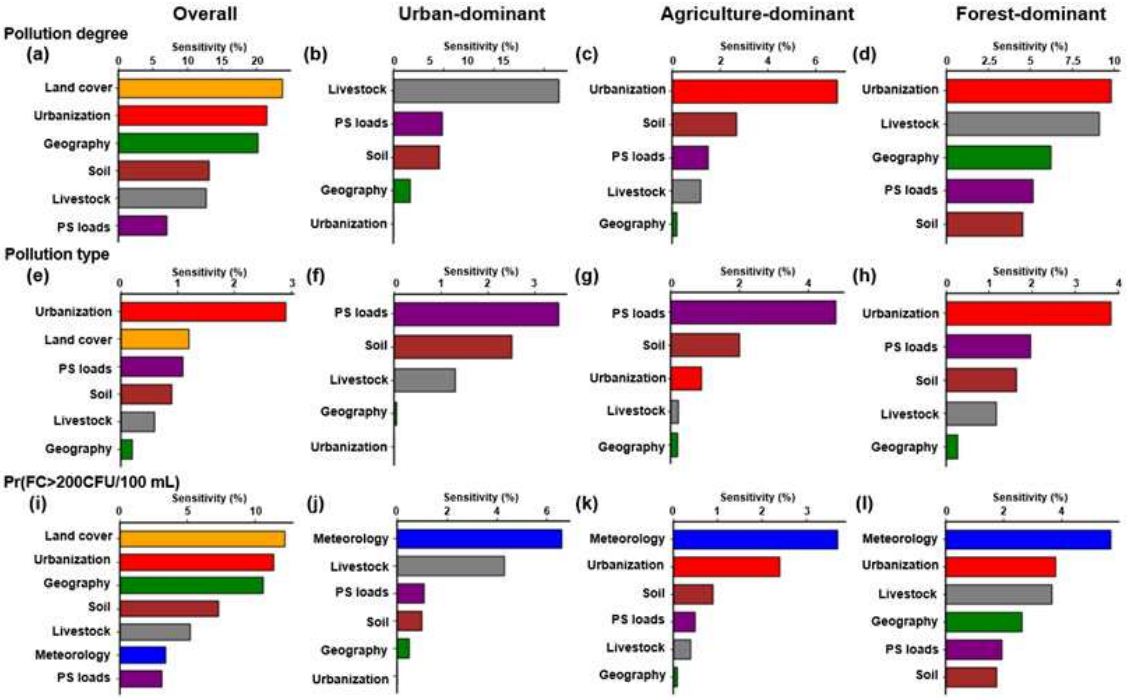
- 베이지안 네트워크는 도식적 모델로 변수 사이의 상호 인과관계를 기반으로 해석 가능한 모델 구조를 제공하므로 모델해석기법의 적용 없이 시나리오 분석 및 민감도 분석 등을 수행하여 영향인자간의 상대적 중요도를 분석한다 <그림 2-40>.



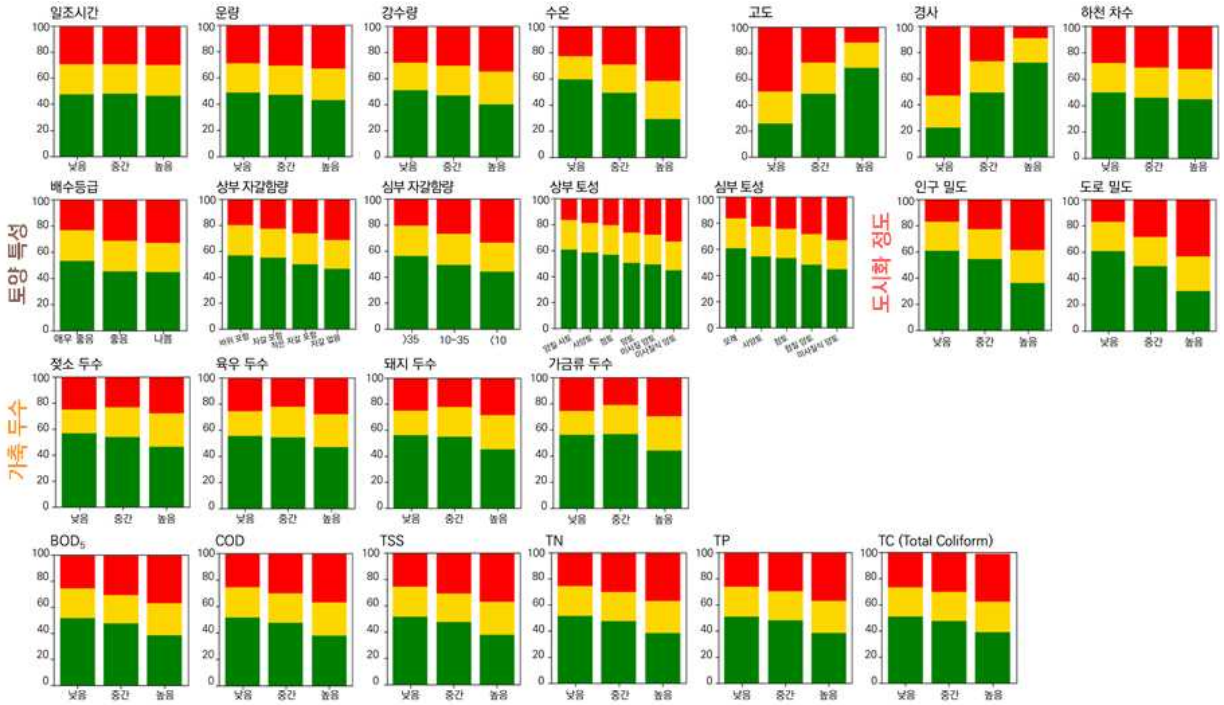
<그림 2-40> 계층적 베이지안 네트워크에 대한 민감도 분석 및 시나리오 분석 적용 예시

✓ **조건부확률 모델을 통한 주요영향인자 분석 예시**

- 해석 가능한 조건부확률 모델(베이지안 네트워크)에 대해서는 민감도 분석 및 시나리오 분석 등을 수행하여 변수의 상대적 중요도 및 변화에 따른 목표기준 달성/미달성 확률 변화 등에 대한 분석을 수행한다 <그림 2-41>, <그림 2-42>.



<그림 2-41> 베이지안 네트워크를 활용한 민감도 분석 예시



<그림 2-42> 베이지안 네트워크를 활용한 시나리오 분석 예시

2-3. 월별 추진 일정

- 월별 추진 일정은 아래와 같다.

<그림 2-43> 월별 추진 일정표

3. 참여 전문인력 및 조직 편성표

3-1. 연구원 구성표

구 분	소 속	직 위	성 명	전 공	담당분야	참여율
연구책임자	서울시립대학교	교수	차윤경	환경공학	연구 총괄	21.5%
책임연구원급	서울시립대학교	교수	이상철	환경공학	결과 검토 및 조정	21.5%
연구원급	-	-	-	-	-	
연구보조원급	서울시립대학교	박사 수료	신지훈	환경공학	모델 개발 및 해석	28%
		박사 재학생	김영우	환경공학	모델 개발 및 해석	28%
		석사 수료	김태호	환경공학	모델 개발	28%
		석사 재학생	김낙겸	환경공학	기초자료 분석	28%
		석사 수료	정혜민	환경공학	모델 개발	28%
		석사 재학생	김동호	환경공학	기초자료 분석	28%
		석사 재학생	권용성	환경공학	기초자료 분석	28%
보조원급	서울시립대학교	학부 재학생	이건형	환경공학	문헌 조사	23%
		학부 재학생	박태승	환경공학	문헌 조사	23%
		학부 재학생	이도연	환경공학	기초자료 수집	23%
		학부 재학생	이지원	환경공학	기초자료 수집	23%

3-2. 연구참여자 인적사항

3-2-1. 연구책임자

3-2-1-1. 인적사항

성 명	차윤경 (한자) 車允璟			생년월일		
소 속	서울시립대학교 환경공학부		직 위	부교수	전 공	환경공학
연락처	사 무 실	02-6490-5504	팩 스	-	전자우편	ykcha@uos.ac.kr
	휴대전화	010-9152-5355				

3-2-1-2. 학력 및 경력사항

	기 간	학 교	전 공	학 위	비 고
학력	1999 년 03 월 - 2004 년 02 월	이화여자대학교	환경공학	학사	
	2004 년 03 월 - 2006 년 02 월	이화여자대학교	환경공학	석사	
	2006 년 03 월 - 2011 년 02 월	미국 듀크대학교	환경학	박사	
	년 월 - 년 월				
	년 월 - 년 월				

	기 간	기 관	직 위	비 고
경력	2011 년 03 월 - 2014 년 02 월	미국 미시건대학교	박사후 연구원	
	2014 년 03 월 - 2015 년 02 월	광주 과학기술원 환경공학과	연구교수	
	2015 년 03 월 - 현재	서울시립대 환경공학부	부교수	
	년 월 - 년 월			
	년 월 - 년 월			

3-2-1-3. 연구실적(최근 3년간)

구분	역할	연구과제명	연구비		연구기간 (부터-까지)	논문발표 학술지명
			금 액	지원기관		
완료	책임연구원	통계적 기법을 활용한 수질변화 평가방안 마련	30,000,000 원	국립환경과학원	2018.10~2019.01	
	참여연구원	소비자중심의 옥내 잔류염소 균등화를 위한 제도개선 및 기술개발	176,400,000 원	한국상하수도협회	2017.03~2019.03	
	책임연구원	녹조 예측을 위한 피코시아닌 고유광특성 산정 알고리즘 개발	150,000,000 원	한국연구재단	2016.06~2019.05	Desalination and Water Treatment, Water Research
	책임연구원	기후변화 대응을 위한 빅데이터 기반의 서울형 하천수질관리시스템 핵심요소기술 개발	117,315,000 원	서울시 보건환경연구원	2018.10~2019.08	

완료	책임연구원	차세대 수질·수생태계 예측모델 개발 및 적용성 평가(II)-국내 수질·수생태계 모니터링 자료를 활용한 수생태모델 활용성 평가	43,956,000 원	국립환경과학원	2018.10~2019.10	
	책임연구원	통계적 기법을 활용한 수질변화 평가방안 마련(II)	99,000,000 원	국립환경과학원	2019.07~2020.03	
	참여연구원	ICT기반 상수도시설 스마트 자산관리 시스템 개발	198,000,000 원	환경부/한국환경산업기술원	2016.08~2021.04	
	책임연구원	다중모드 환경 데이터의 다중규모 분석을 위한 인공지능 기반 모듈화 모델 개발	100,000,000 원	한국연구재단 (1차년도)	2020.03~2020.12	
	책임연구원	상하수도 빅데이터 관리·해석 플랫폼 및 표준화 기술 개발	55,666,400 원	환경부/한국환경산업기술원 (1차년도)	2020.03~2020.12	
	참여연구원	정수생태계 건강성평가 방법 및 변화예측 모델 개발	40,028,000 원	환경부/한국환경산업기술원 (1차년도)	2020.03~2020.12	
	책임연구원	서울 미세먼지 대응 정책 지원을 위한 서울시 도로변 대기오염측정망 최적화 기법 개발	30,000,000 원	서울녹색환경지원센터	2020.04~2020.12	
수행 중인 과제	책임연구원	상하수도 빅데이터 관리·해석 플랫폼 및 표준화 기술 개발	88,000,000 원	환경부/한국환경산업기술원 (2차년도)	2021.01~2021.12	
	참여연구원	정수생태계 건강성평가 방법 및 변화예측 모델 개발	52,000,000 원	환경부/한국환경산업기술원 (2차년도)	2021.01~2021.12	
	책임연구원	다중모드 환경 데이터의 다중규모 분석을 위한 인공지능 기반 모듈화 모델 개발	100,000,000 원	한국연구재단 (2차년도)	2021.01~2021.12	

3-2-1-4. 저서실적

발행년도	저 서 명	출 판 사	발 행 지 (국내, 국외)	비 고

3-2-1-5. 연구논문 발표실적(최근 3년간)

연구과제 명	연구기간 (부터 - 까지)	연구논문 발표지명 (발표년월)	역 할 (책임자, 연구원)	연구비 지원기관
The Effects of Tree Species on Soil Organic Carbon Content in South Korea		Journal of Geophysical Research-Biogeosciences (2019.03)	연구원	한국연구재단
The implicaitons of Simpson’s paradox for cross-scale inference among lakes		Water Research (2019.10)	연구원	Great Lakes Environmental Research Laboratory of the US National Oceanic and Atmospheric Administration
Evaluation of temperature effects on brake wear particles using clustered heatmaps		Environmental Engineering Research (2019.12)	연구원	한국연구재단
Simulating seasonal variability of phytopalnkton in stream water using the modified SWAT model		Environmental Modelling and Software (2019.12)	연구원	한국연구재단
Deep learning-based retrieval of cyanobacteria pigment in inland water for in-situ and airtborne hyperspectral data		Ecological Indicators (2020.03)	책임자	환경부
An Integrative Remote Sensing Application of Stacked Autoencoder for Atmospheric Correction and Cyoanobacteria Estimation Using Hyperspectral Imagery		REMOTE SENSING (2020.04)	연구원	환경부, 서울시립대학교
Assessing Land-Cover Effects on Stream Water Quality in Metropolitan Areas Using the Water Quality Index		Water (2020.11)	책임자	한국연구재단
Effects of class imbalance on resampling and ensemble learning for improved predictions of cyanobacteria bloom		Ecological Informatics (2021.03)	연구원	한국환경산업기술원
Forecasting abrupt depletion of dissolved oxygen in urban streams using discontinuously measured hourly timeseries data		Water Resources Research (2021.03)	책임자	한국연구재단
Willingness to Pay for Improved Water Supply Services Based on Asset Management: A Contingent Valuation Study in South Korea		Water (2021.07)	연구원	한국환경산업기술원
An interpretable machine learning method for supporting ecosystem management: Application to species distribution models of freshwater macroinvertebrates		Journal of Environmental Management (2021.08)	책임자	한국환경산업기술원 한국연구재단
자료기반 물환경 모델의 현황 및 발전 방향		한국물환경학회지 (2020.11)	책임자	한국환경산업기술원
수질지수와 군집분석을 활용한 서울시 주요 하천 수질평가		대한환경공학회지 (2020.11)	연구원	서울특별시
베이지안 최적화를 통한 저서성 대형무척추동물 종분포모델 개발		대한상하수도학회지 (2021.08)		한국환경산업기술원

3-2-2. 책임 연구원급

3-2-2-1. 인적사항

성 명	이상철 (한자) 李相澈			생년월일	84.03.19	
소 속	서울시립대 환경공학부		직 위	조교수	전 공	빅데이터
연락처	사 무 실	02-6490-2875	팩 스		전자 우편	slee2020@uos.ac.kr
	휴대전화	010-8301-3419				

3-2-2-2. 학력 및 경력사항

	기 간	학 교	전 공	학 위	비 고
학 력	2003년 3월 - 2007년 2월	고려대학교	환경생태	학사	
	2009년 3월 - 2011년 8월	고려대학교	환경생태	석사	
	2011년 8월 - 2017년 5월	미국 메릴랜드 주립대학교	Geographical Sciences	박사	
	년 월 - 년 월				

	기 간	기 관	직 위	비 고
경 력	2013년 5월 - 2017년 2월	미농무성(USDA)	방문학생	
	2017년 5월 - 2020년 8월	미농무성(USDA)	방문연구원	
	2020년 9월 - 현재	서울시립대 환경공학부	조교수	
	년 월 - 년 월			

3-2-2-3. 연구실적 (최근 3년간)

구분	역할	연구과제명	연구비		연구기간 (부터-까지)	논문발표 학술지명
			금 액	지원기관		
완료						
수행 중인 과제	책임연구원	빅데이터-인공지능 기반 스마트 물관측체계 개발	121,155,204	한국연구재단 (1차년도)	2021.03~2022.02	
	참여연구원	지중 내 유기오염물질의 선택적 차단을 위한 스마트 차수재 개발 및 AI기법을 이용한 환경영향 예측 평가	200,000,000	환경부/한국환경산업기술원 (1차년도)	2021.07~2021.12	
	참여연구원	수생태계 물질 순환에 기반한 생태계 서식환경 예측 및 그린 인프라 통합 관리 시스템 개발	45,000,000	한국연구재단 (1차년도)	2021.09~2022.02	

3-2-2-4. 저서실적 (최근 3년간)

발행년도	저 서 명	출 판 사	발 행 지 (국내, 국외)	비 고

3-2-2-5. 연구논문 발표실적(최근 3년간)

연구과제명	연 구 기 간 (부터 - 까지)	연구논문 발표지명 (발표년월)	역 할 (책임자, 연구원)	연구비 지원기관
Assessing the cumulative impacts of geographically isolated wetlands on watershed hydrology using the SWAT model coupled with improved wetland modules		Journal of Environmental Management (2018.10)	책임자	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Effects of subsurface soil characteristics on wetland-groundwater interaction in the coastal plain of the Chesapeake Bay watershed		Hydrological Processes (2019.01)	책임자	U.S. Department of Agriculture (USDA)
Mapping the landscape-level hydrologic connectivity of headwater wetlands to downstream water: a geospatial modelling approach – Part I		Science of the Total Environment (2019.02)	연구원	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Mapping the landscape-level hydrologic connectivity of headwater wetlands to downstream water: a geospatial modelling approach – Part 2		Science of the Total Environment (2019.02)	연구원	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Enhancement of Agricultural Policy/Environment eXtender (APEX) Model to Assess Effectiveness of Wetland Water Quality Functions		Water (2019.03)	책임자	U.S. Department of Agriculture (USDA)
IPEAT+: A Built-In Optimization and Automatic Calibration Tool of SWAT+		Water (2019.08)	연구원	U.S. Department of Agriculture (USDA)
A coupled surface water storage and subsurface water dynamics model in SWAT for characterizing hydroperiod of geographically isolated wetlands		Advances in Water Resources (2019.09)	연구원	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Improving the catchment scale wetland modeling using remotely sensed data		Environmental Modelling & Software (2019.12)	책임자	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Evaluation of the Soil Vulnerability Index for artificially drained cropland across eight Conservation Effects Assessment Project watersheds		Journal of Soil and Water Conservation (2020.01)	연구원	U.S. Department of Agriculture (USDA)
Mapping Forested Wetland Inundation in the Delmarva Peninsula, USA Using Deep Convolutional Neural Networks		Remote Sensing (2020.02)	연구원	U.S. Department of Agriculture (USDA)
Modeling riverine dissolved and particulate organic carbon fluxes from two small watersheds in the northeastern United States		Environmental Modelling & Software (2020.02)	연구원	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Effects of surface runoff and infiltration partition methods on hydrological modeling: A comparison of four schemes in two watersheds in the Northeastern US		Journal of Hydrology (2020.02)	연구원	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)

Use of multiple modules and Bayesian Model Averaging to assess structural uncertainty of catchment-scale wetland modeling in a Coastal Plain landscape		Journal of Hydrology (2020.03)	책임자	U.S. Department of Agriculture (USDA)
Seasonal drivers of geographically isolated wetland hydrology in a low-gradient, Coastal Plain landscape		Journal of Hydrology (2020.04)	책임자	U.S. Department of Agriculture (USDA)
Assessing the effectiveness of riparian buffers for reducing organic nitrogen loads in the Coastal Plain of the Chesapeake Bay watershed using a watershed model		Journal of Hydrology (2020.06)	책임자	U.S. Department of Agriculture (USDA)
Estimating the effect of winter cover crops on nitrogen leaching using cost-share enrollment data, satellite remote sensing, and Soil and Water Assessment Tool (SWAT) modeling		Journal of Soil and Water Conservation (2020.05-06)	연구원	U.S. Department of Agriculture (USDA)
Use of Topographic Models for Mapping Soil Properties and Processes		Soil Systems (2020.06)	연구원	U.S. Department of Agriculture (USDA)
Modeling sediment diagenesis processes on riverbed to better quantify aquatic carbon fluxes and stocks in a small watershed of the Mid-Atlantic region		Carbon Balance and Management (2020.07)	연구원	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Evaluating a remote wetland functional assessment along an alternation gradient in coastal plain depressional wetlands		Journal of Soil and Water Conservation (2020.11-12)	연구원	U.S. Department of Agriculture (USDA)
Overview of the USDA Mid-Atlantic Regional Wetland Conservation Effects Assessment Project		Journal of Soil and Water Conservation (2020.11-12)	책임자	U.S. Department of Agriculture (USDA)
Assessment and Combination of SMAP and Sentinel-1A/B-Derived Soil Moisture Estimates With Land Surface Model Outputs in the Mid-Atlantic Coastal Plain, USA		IEEE Transactions on Geoscience and Remote Sensing (2021.02)	책임자	U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Utility of remotely sensed evapotranspiration products on assessing an improved model structure		Sustainability (2021.02)	책임자	U.S. Department of Agriculture (USDA)
Uncertainty assessment of multi-parameter, multi-GCM, and multi-RCP simulations for streamflow and non-floodplain wetland (NFW) water storage		Journal of Hydrology (2021.06)	책임자	한국연구재단, U.S. Department of Agriculture (USDA) and National Aeronautics and Space Administration (NASA)
Spatial extrapolation of topographic models for mapping soil organic carbon using local samples		Geoderma (2021.06)	연구원	한국연구재단, U.S. Department of Agriculture (USDA)

3-2-3. 연구 보조원급

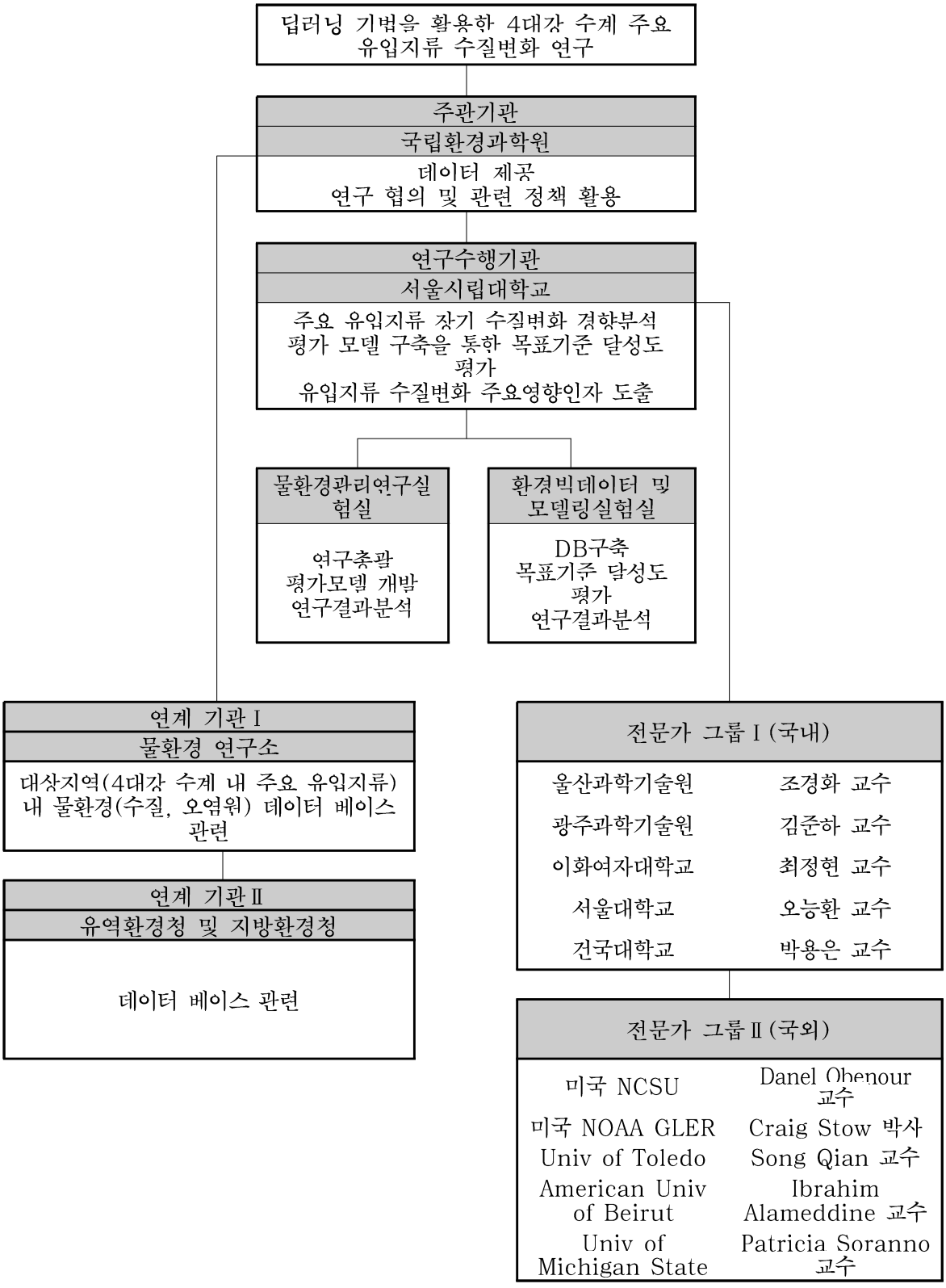
소 속	직 위	성 명	생년 월일	최종출신 학 교	학위	전공	연락처	담당분야
서울시립대	학생연구원	신지훈	1993.01.04	서울시립대	박사수료	환경공학	010-2288-7839	모델 개발 및 해석
서울시립대	학생연구원	김영우	1992.05.08	서울시립대	박사재학	환경공학	010-7236-1524	모델 개발 및 해석
서울시립대	학생연구원	김태호	1995.02.16	서울시립대	석사수료	환경공학	010-4616-2438	모델 개발
서울시립대	학생연구원	김낙겸	1995.04.21	서울시립대	석사재학	환경공학	010-4915-7616	기초자료 분석
서울시립대	학생연구원	정혜민	1993.11.18	서울시립대	석사수료	환경공학	010-2084-5398	모델 개발
서울시립대	학생연구원	김동호	1994.09.21	서울시립대	석사재학	환경공학	010-4074-8749	기초자료 분석
서울시립대	학생연구원	권용성	1995.04.10	서울시립대	석사재학	환경공학	010-4440-5873	기초자료 분석

3-2-4. 보조원급

소 속	직 위	성 명	생년 월일	최종출신 학 교	학위	전공	연락처	담당분야
서울시립대	학생연구원	박태승	1997.10.03	서울시립대	학사재학	환경공학	010-2261-4525	문헌 조사
서울시립대	학생연구원	이건형	1997.08.18	서울시립대	학사재학	환경공학	010-8381-3433	문헌 조사
서울시립대	학생연구원	이도연	1999.11.12	서울시립대	학사재학	환경공학	010-5502-6387	기초자료 수집
서울시립대	학생연구원	이지원	1997.11.04	서울시립대	학사재학	환경공학	010-4055-2888	기초자료 수집

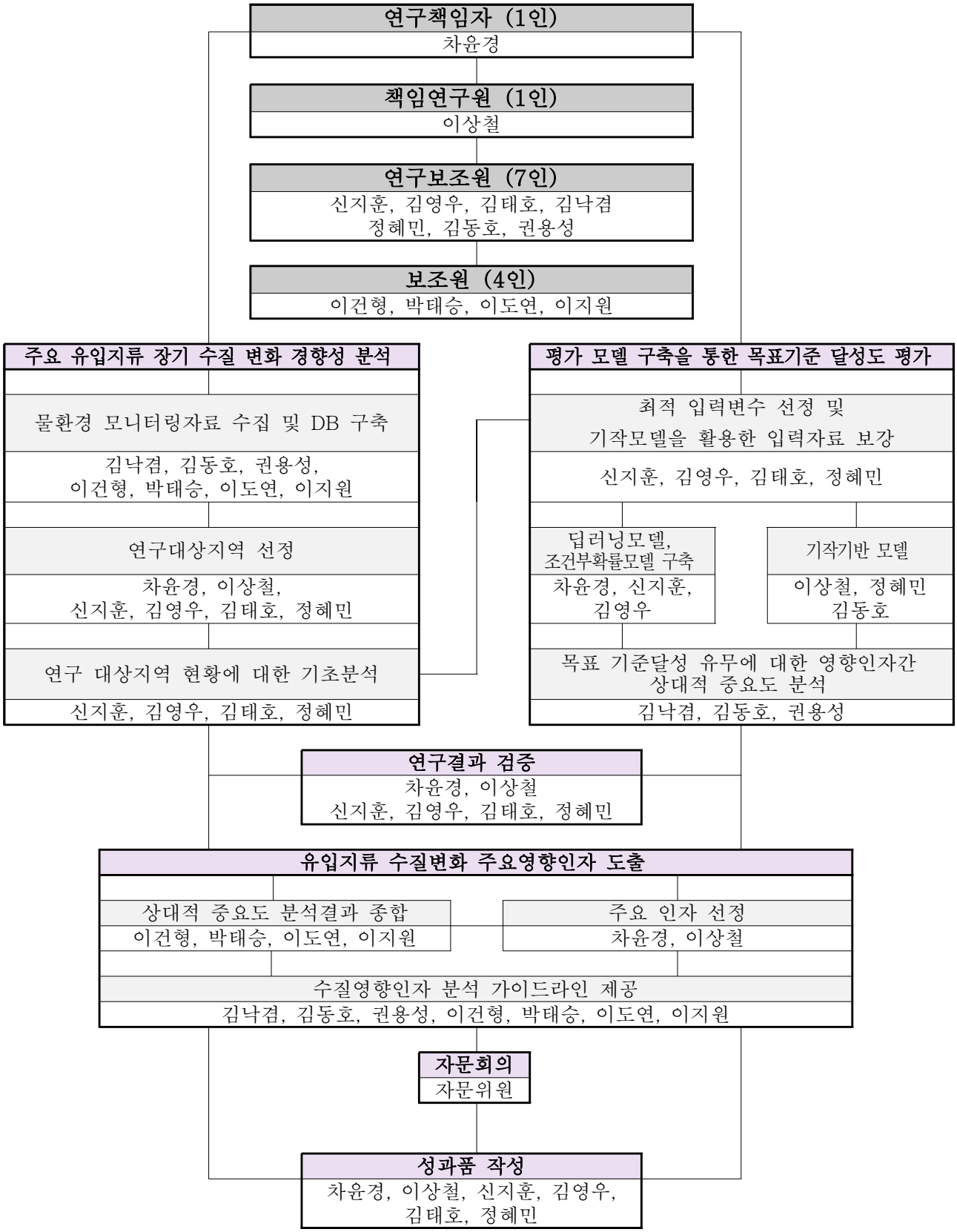
3-3. 연구 수행 조직도

3-3-1. 연구기관 간 역할 및 연구 자문 체계



<그림 3-1> 연구기관 역할 및 연구자문 체계

3-3-2. 과업수행 체계 및 조직 편성



<그림 3-2> 과업수행 체계

4. 산출내역서

최종 산출내역서

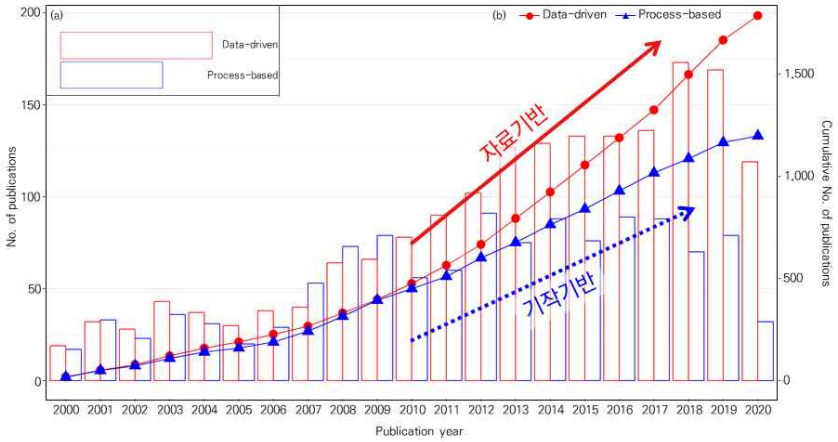
(단위 : 원, %)

비목 / 구분	금 액	구성비	산 출 근 거
1.인 건 비	92,475,660	73.4	
책 임 연 구 원 급	22,331,640	17.7	6,433,726원(단가)×1인×8개월×21.5%(참여율)
연 구 원 급	-		-
연 구 보 조 원 급	52,174,980	41.4	3,297,486원(단가)×4인×8개월×28%(참여율)
보 조 원 급	17,969,040	14.3	2,495,700원(단가)×3인×8개월×23%(참여율)
2.경 비	15,243,560	12.1	
여 비	3,300,000	2.6	책임연구원급 300,000원(단가)×2인×1회(2박3일) 연구원급이하 170,000원(단가)×7인×2회(1박2일) 연구원급이하 40,000원(단가)×4인×2회(관외비숙박)
유 인 물 비	5,013,560	4.0	중간보고서 435,620원 최종보고서 780,120원 요약보고자료 297,820원 위탁정산수수료 500,000원 논문교정비 3,000,000원
전 산 처 리 비	2,500,000	2.0	200,000원(토너단가)×10개 25,000원(복사용지)×20개
시 약 및 재 료 비	1,600,000	1.3	200,000원(사무용품비)×8개
회 의 비	1,800,000 150,000	1.5	자문회의 수당 150,000원(단가)×6인×2회 회의준비비 5,000원×6인×5회
임 차 료	450,000	0.4	회의실임대비 150,000원×3회
교 통 통 신 비	430,000	0.3	20,000원(시내교통비)×4인×5회 10,000원(우편요금)×3회
3.일 반 관 리 비(%)	6,826,235	5.4	인건비 + 경비의 6% 이내
4.이 윤(%)	-		-
5.위 탁 연 구 개 발 비	-		-
6.부 가 가 치 세(%)	11,454,545	9.1	인건비+경비+일반관리비의 10%
7.총 연 구 비	126,000,000	100	

5. 기타 관련 연구자료

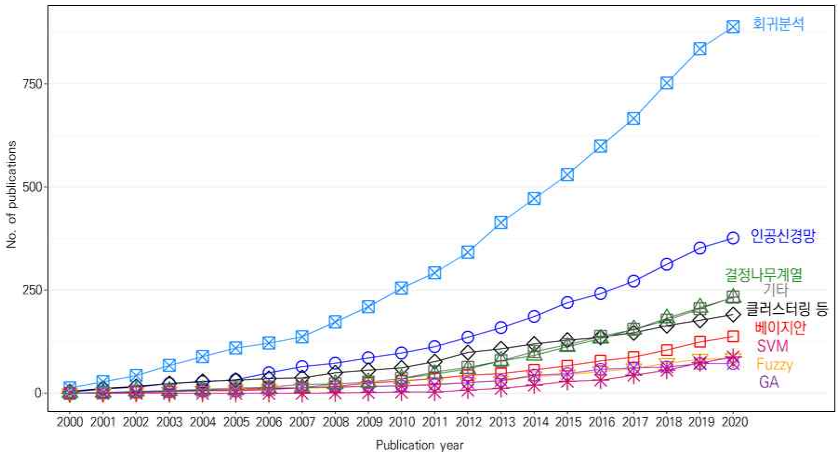
5-1. 국내·외 연구 동향

- 국제 학술 데이터베이스인 Web of Science 활용해 최근 20년간 (2000년-2020년)의 물환경 모델링 관련 문헌을 조사한 결과, 선정된 2,944편의 문헌 중 자료기반 모델이 총 1,784편, 기작기반 모델이 총 1,200편의 문헌에서 활용된 것을 확인할 수 있었다.
- 2010년을 기점으로 자료기반 모델링 기법의 활용 증가율이 기작기반 모델보다 높아짐을 확인할 수 있었다 <그림 5-1>.



<그림 5-1> 자료기반 및 기작기반 모델 활용 문헌의 출판 현황(차윤경 외 2019)

- 자료기반 모델에 대한 활용 기법을 살펴보면, 가장 기본적인 자료기반 모델인 회귀모델을 제외하면, 딥러닝을 포함하는 인공신경망 계열 모델의 활용이 타 자료기반 모델링보다 눈에 띄게 많았으며, 그 다음으로는 결정나무 기반 모델, 클러스터링 및 차원축소 기법, 베이지안 모델링의 순으로 활용사례가 많았다 <그림 5-2>.



<그림 5-2> 자료기반 모델의 세분류별 활용 문헌의 출판 현황(차윤경 외 2019)

- 자료기반 모델의 활용 증가와 함께 수질예측 및 수질변화 분석에 있어 기작기반 모델과 자료기반 모델을 융합하여 각 방법론의 장, 단점을 상호보완하는 시도를 확인할 수 있었다.
- 자료기반 모델과 기작기반 모델의 융합은 기작기반 모델의 파라미터 추정에 자료기반 모델을 활용한 사례, 기작기반 모델을 통합모델의 모듈로 활용한 사례, 기작기반 요소를 자료기반 모델의 구조에 반영한 사례 등이 있었다.
  - 기작기반 모델의 파라미터 추정에 자료기반 모델을 사용한 사례에는 대표적으로 베이지안 프레임워크를 활용하여 기작기반 모델의 파라미터를 추정, 보정(calibartion)을 수행하고 구축된 모델을 바탕으로 호소의 부영양화를 분석한 사례가 있다(Arhonditsis et al., 2007).
  - 기작기반 모델을 자료기반 통합모델의 모듈로 활용한 사례에는 대표적으로 다양한 기작기반 모델과 베이지안 네트워크를 결합한 통합모델을 통해 토지이용, 기후변화 등에 따른 수질 및 수생태계 변화를 모의한 사례가 있다(Couture et al., 2018).
  - 기작기반 요소를 자료기반 모델의 구조에 반영한 사례에는 대표적으로 SPARROW 모델을 활용하여 지역(regional) 단위의 질소 흐름(nitrogen flux) 변화를 모델링 한 사례가 있다 (Hoos et al., 2019; Alexander et al., 2000).
- 최근에는 이에 더해 자료기반 모델의 손실함수에 대한 가이드(guide)에 기작기반 지식이나 원리를 활용한 사례를 확인할 수 있었다 <표 5-1>.

<표 5-1> 기작기반 모델을 자료기반 모델 손실함수의 가이드로 활용한 사례

저자	게재 년도	학술지명	연구 대상지	종속변수	기작 모델	자료 모델
Read et. al.	‘19	Water Resources Research	Lakes in the Minnesota and Wisconsin	Water temperature	GLM	RNN
Jia et. al.	‘20	arXiv	Delaware River Basin	Stream flow, Water temperature	PRMS-SNTemp	RGrN
Hanson et. al.	‘20	Ecological Modelling	Lake Mendota	Eplilimnetic P	Mass balance model	RNN
Noori et. al.	‘20	Journal of Hydrology	City of Atlanta	Nutrient Loads	SWAT	ANN

5-1-1. 국내・외 인공신경망 활용연구 동향

- 국외 문헌조사 결과, 다양한 인공신경망 모델이 물환경 분야에서 활용됨을 확인할 수 있었다.
  - 미국에서는 클로로필a(Milie et al., 2006), Secchi Disk Depth(Heddam, 2016) 등과 같은 수질항목의 예측에 대해 인공신경망(artificial neural network)을 활용한 사례가 있다.
  - 또한, LSTM(long short-term memory)를 활용하여 DO를 예측한 사례가 있다.
  - 캐나다에서는 기후변화 시나리오를 활용, 대상지역의 강수량 변화에 따른 수질 예측에 인공신경망 모델을 활용한 사례가 있다(He et al., 2011).
  - 중국에서는 하천의 상・하류에서 pH, 과망간산염, T-N, DO 등 수질 변화의 예측에 순환신경망(recurrent neural network)을 활용한 사례가 있다(Li et al., 2019).
  - 인도에서는 하천의 DO와 BOD의 예측을 위해 인공신경망 모델을 활용한 사례가 있으며(Basant et al., 2010), Bi-LSTM 모델을 활용하여 수체 BOD의 변화를 예측한 사례가 있다(Khullar and Singh, 2021).
  - 그 외 국가에서도 수질예측을 위해 인공신경망 모델로 하천 상・하류의 전기전도도와 총용존고형물을 예측하거나(Shah et al., 2021), 기상데이터를 입력변수로 수온을 예측한 사례가 있다(Temizyure et al., 2018).
  - 이 외에도 다양한 연구들이 활발히 진행 중에 있다.
- 국내 문헌조사 결과, 다양한 딥러닝 기법들이 물환경 분야에서 사용됨을 확인할 수 있었다.
  - 한강 유역에서는 순환신경망 모델을 이용하여 유해남조류세포수를 (Hong et al., 2020), LM(Levenberg-Marquardt)와 MD(modula) 신경망 모델로 BOD, TN, TP를 예측하였다(Cho et al., 2004).
  - 또한, 남한강 유역에서 LMNN, ANFIS, MDNN 모델을 이용하여 DO, TOC, TN, TP를 수체내 변화를 확인한 사례가 있다(Yeon et al., 2005).
  - 낙동강 유역에서는 BOD와 DO의 예측(Cho., 2000), 클로로필a의 예측(Park et. al., 2014)을 위해 인공신경망을 활용한 사례가 있다.
  - 최근에는 순환신경망 및 LSTM 모델을 이용하여 낙동강 온천천의 DO를 예측한 사례가 있었다(Lim et al., 2020).

- 금강 유역에서는 DO, BOD, T-N의 예측을 위해 모멘트-적응학습율 방법을 적용한 MANN(Moment-Adaptive learning rate neural network), Lavenberg-Marquardt 방법을 이용한 LMNN(Levenberg-Marquardt neural network), 은닉층을 분리한 MNN(Modular Neural Network)를 활용한 뒤 각 모델을 비교한 사례가 있다(Ahn et al., 2001).
- 영산강 유역에서는 TOC의 예측을 위해 종속변수에 대한 웨이블렛(wavelet) 변환을 적용한 후 인공신경망을 활용, 향상된 예측 정확도를 확인한 사례가 있다(Oh et al., 2008).
- 섬진강 유역에서는 BOD, CO, SS의 예측을 위해 인공신경망을 활용한 사례가 있다(Kim et al., 2001).

- 위 사례는 5대강 수계를 대상으로 이루어진 국내 학술지 출판물로, 국제학술지에도 국내 연구 대상지에 관한 다양한 연구들이 수행되고 있다.
- 국내・외 문헌조사 결과, 딥러닝 등 인공신경망 계열 모델을 통한 수질항목의 예측에 관한 연구는 활발히 이루어지고 있으나, 상대적으로 예측 결과에 대한 해석력 확보를 위한 연구는 부족한 실정이다.
- 그러나, 예측결과에 대한 해석력이 뒷받침되지 않는다면 현상이나 시스템에 대한 새로운 통찰이나 이해를 얻기에는 어려움이 있으므로 해석력 확보를 위한 연구가 필요하다.

5-1-2. 국내・외 베이지안 네트워크 활용연구 동향

- 국외 문헌조사 결과, 물환경 분야에서 베이지안 네트워크가 활용되고 있음을 확인할 수 있었다.
  - 미국에서는 Neuse River에 대한 부영양화의 영향 분석을 위해 베이지안 네트워크(Bayesian network)를 적용한 사례(Borsuk et al., 2004), Ohio 주 내의 하천에서 수질 및 수리・수문 변화에 따른 저서성 대형무척추동물 중 수질 지표종인 EPT taxa에 대한 적용사례(McLaughlin and Reckhow, 2017) 등이 있다.
  - 호주에서는 Upper Murrumbidgee River에서 대형 무척추동물의 상대적 분포의 예측을 위해 적용한 사례(Lucena-Moya, 2015), 다양한 호수에서 수질 등의 변화에 따른 녹조현상의 위험도 분석에 적용한 사례(Rigosi et al., 2015), Tasmania의 지류에서 서식지 적합성을 비교한 사례(Hamilton et al., 2015) 등이 있다.

- 유럽에서는 다양한 연못에서 생태, 환경변수 등에 따른 생태계 서비스(ecosystem service)와 관리 비용(management cost)의 변화 분석을 위해 활용된 사례 등이 있다 (Landuyt et al., 2014).
- 중국에서는 Taizi River 유역에서 토지이용, 수질, 지리변수 등을 활용하여 EPT taxa와 관련된 지표의 변화를 분석한 사례가 있다(Li et al., 2018).

- 국내 문헌조사 결과, 국내에서는 상대적으로 물환경 분야에서 베이지안 네트워크의 활용사례가 많지 않으며 주로 가뭄 관련 연구가 이루어짐을 확인했다 (Shin et al., 2017; Yoo et al., 2014).
- 베이지안 네트워크는 해석적인 측면에서 우수할 뿐만 아니라, 변수 간의 상호 관계 혹은 예측에 대한 불확실성에 대해 조건적인 의존 관계를 확률적으로 표현하므로 환경 정책 등 의사결정을 지원에 있어 매우 적절한 방법으로 수질분석을 위한 적용 연구가 필요하다.

5-1-3. 국내・외 SWAT 모델 활용연구 동향

- 국외 문헌조사 결과, SWAT모델이 수체의 수질변화 분석에 널리 활용됨을 확인할 수 있었다.
  - 미국에서는 SWAT모델로 TN, TP를 예측하고 다른 수문모델을 동시에 적용한 앙상블 기법으로 오염물질을 예측 정확도를 평가했다(Sharifi et al., 2006).
  - 또한 SWAT 내 모듈을 고도화한 SWAT-C를 활용하여 DOC와 POC를 예측하였다 (Qi et al., 2020).
  - 캐나다에서는 SWAT을 고도화하여 클로라이드(chloride) 농도를 예측하고 고도화된 모델의 적용성을 평가했다(Zhang et al., 2020).
  - 유럽에서는 SWAT을 유럽전체 대상지에 적용하여 검보정하여 수량과 질산염 농도를 예측하여 광범위한 지역에서 물환경관리 방안을 모색했다(Abbaspour et al., 2015)
  - 중국 동부에서는 댐과 같은 인공지형물이 수질에 미치는 영향을 평가하기 위해 SWAT을 고도화하여 NH_3-N과 COD_{Mn}의 농도를 예측하였다(Zhang et al., 2011).
  - 호주에서는 TSS, TN, TP의 예측을 위해 여러 지역에 측정한 값을 SWAT의 검・보정 자료로 활용하여 오염물질 농도를 예측하였다(Shrestha et al., 2016).

- 국내 문헌조사 결과, 국내에서도 SWAT모델이 수체의 수질변화 분석에 널리 활용됨을 확인할 수 있었다.
  - 한강 유역에서는 한강수계 내 충주댐 유역에서는 SWAT을 이용하여 Sediments, TN, TP를 예측하였고 최적관리기법(Best Management Practices, BMP)이 Sediments, TN, TP이 저감에 미치는 영향을 평가하였다(Yu et al., 2012).
  - 또한, 만대천 유엑에서는 비점오염원에 배출되는 TP의 저감을 위한 초생대 수질개선 효과를 SWAT으로 예측하였다(Lee et al., 2011).
  - 낙동강 유역을 대상으로 SWAT을 적용하여 본류 및 지류에서 장기간 유출되는 유사량을 예측하였다(Ji et al., 2014).
  - 금강 유역 내 예당 저수지 유역에서 측정된 TN, TP를 대상으로 SWAT을 검보정하고 SWAT-APEX 모델을 구동하여 필드 단위에서의 오염물질 거동을 평가하였다(Jung et al., 2011).
  - 또한, 용담댐 유역에서는 SWAT을 적용하여 기후변화 영향을 예측하는 연구를 수행하였다(Jung et al., 2015).
  - 섬진강 상류에 위치한 요천 유역에서는 SWAT을 적용하여 TSS, TN, TP를 검보정하고 기후변화자료의 보정된 SWAT에 입력하여 기후변화에 따른 오염물질농도 변화를 예측하였다(Jang et al., 2018).
- 국내・외 문헌조사 결과, SWAT모델은 모니터링이 어려운 수체의 수질예측에 효과적인 수단으로 그 이용이 중요해지고 있고, 최근 기계학습/딥러닝을 포함한 인공지능 기술이 SWAT에 적용되어 향상된 결과를 나타낸 사례가 증가하고 있다.
- 그러나 수질예측을 위한 SWAT-인공지능 융복합 연구는 아직 부족한 실정이며, 특히, 국내는 유량 측정소와 비교하여 수질 측정망 지점과 모니터링 자료가 부족한 상황이다. 따라서 SWAT으로 모의된 수질자료를 인공지능에 접목하여 모니터링 사각지대를 최소화하는 연구가 필요하다.