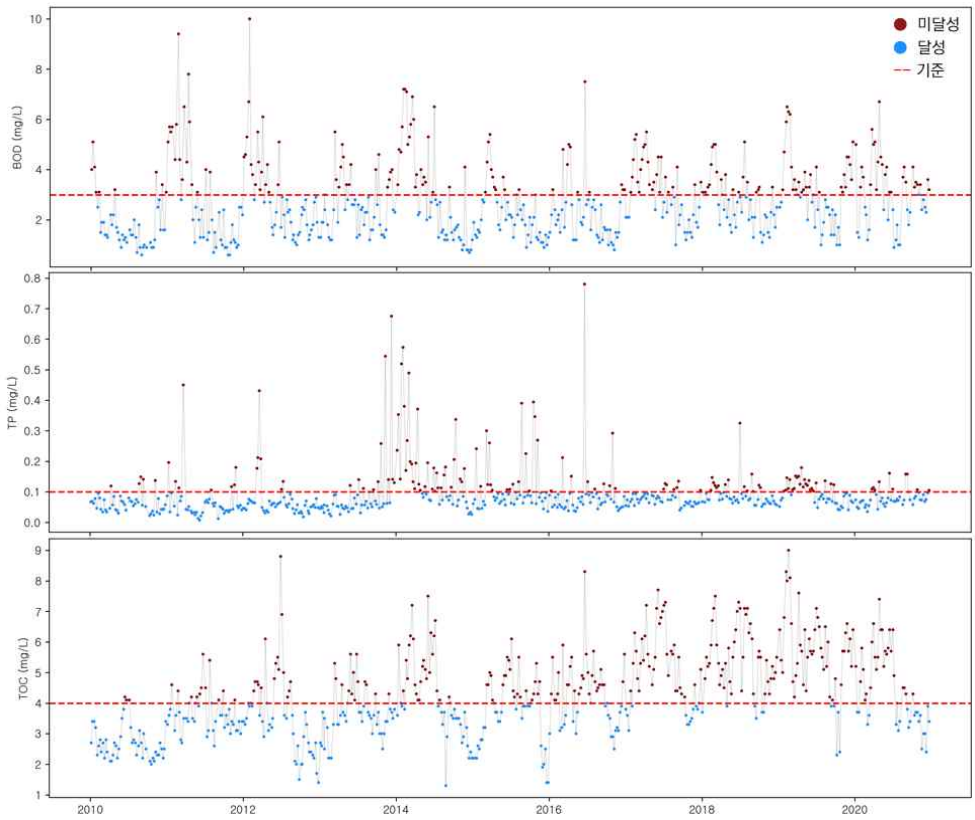
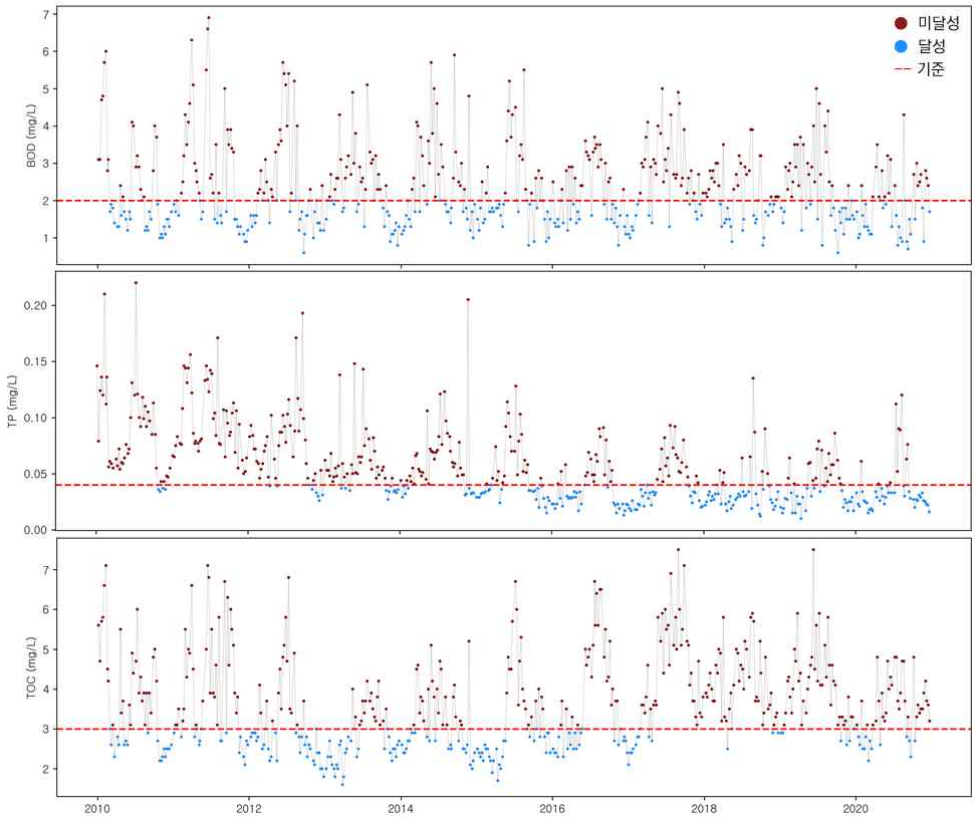


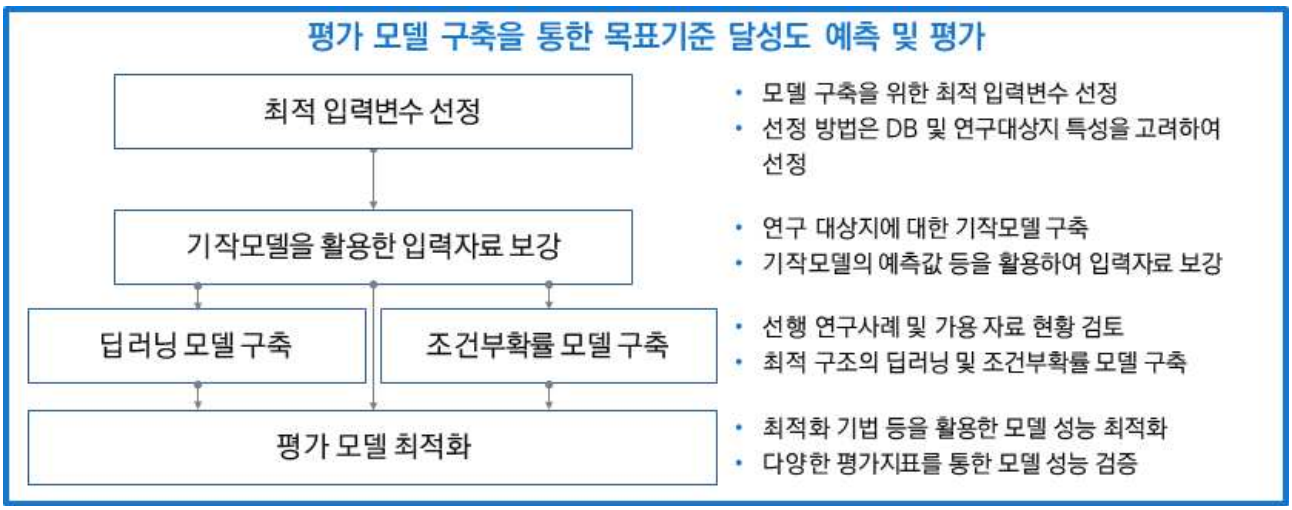
<그림 2-14> 2010-2020 기간 황룡강3-1 지점 BOD, TOC, T-P 농도 변화



<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.

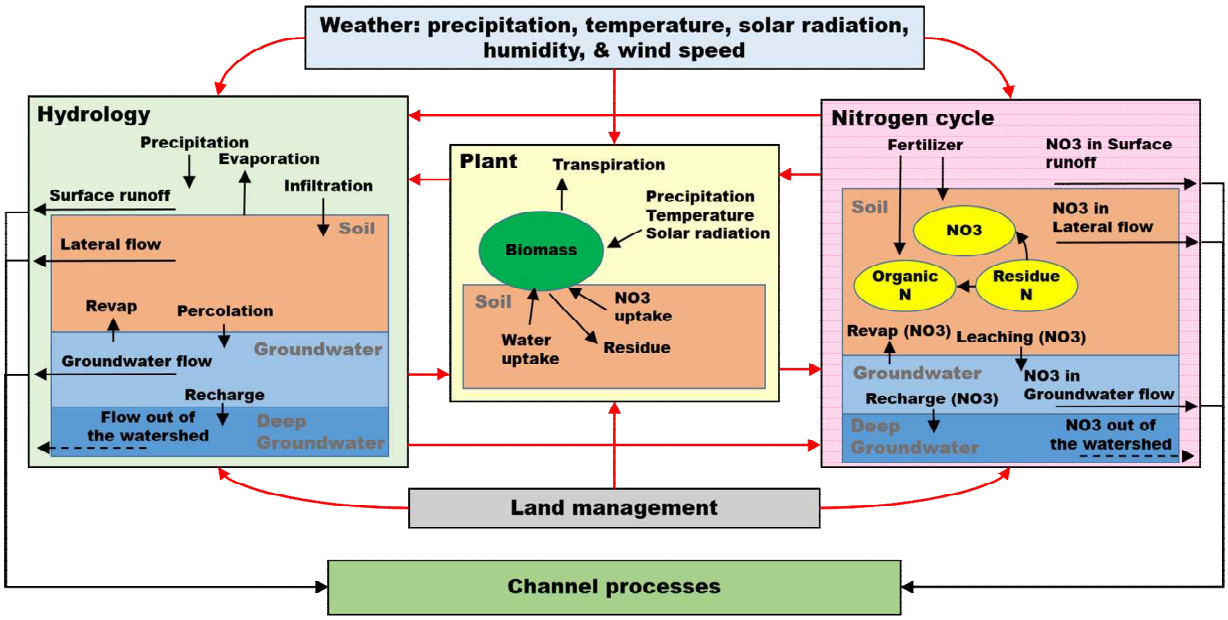
- 마지막으로 주요 수질 항목의 선정 단계에서는 각 요인들에서 요인적재량의 절대값이 가장 큰 항목을 그 요인의 특성을 대표하는 주요 특성으로 판단하고 그 항목을 그 요인의 특성을 대표하는 환경 변수로 판단한다.

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.

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

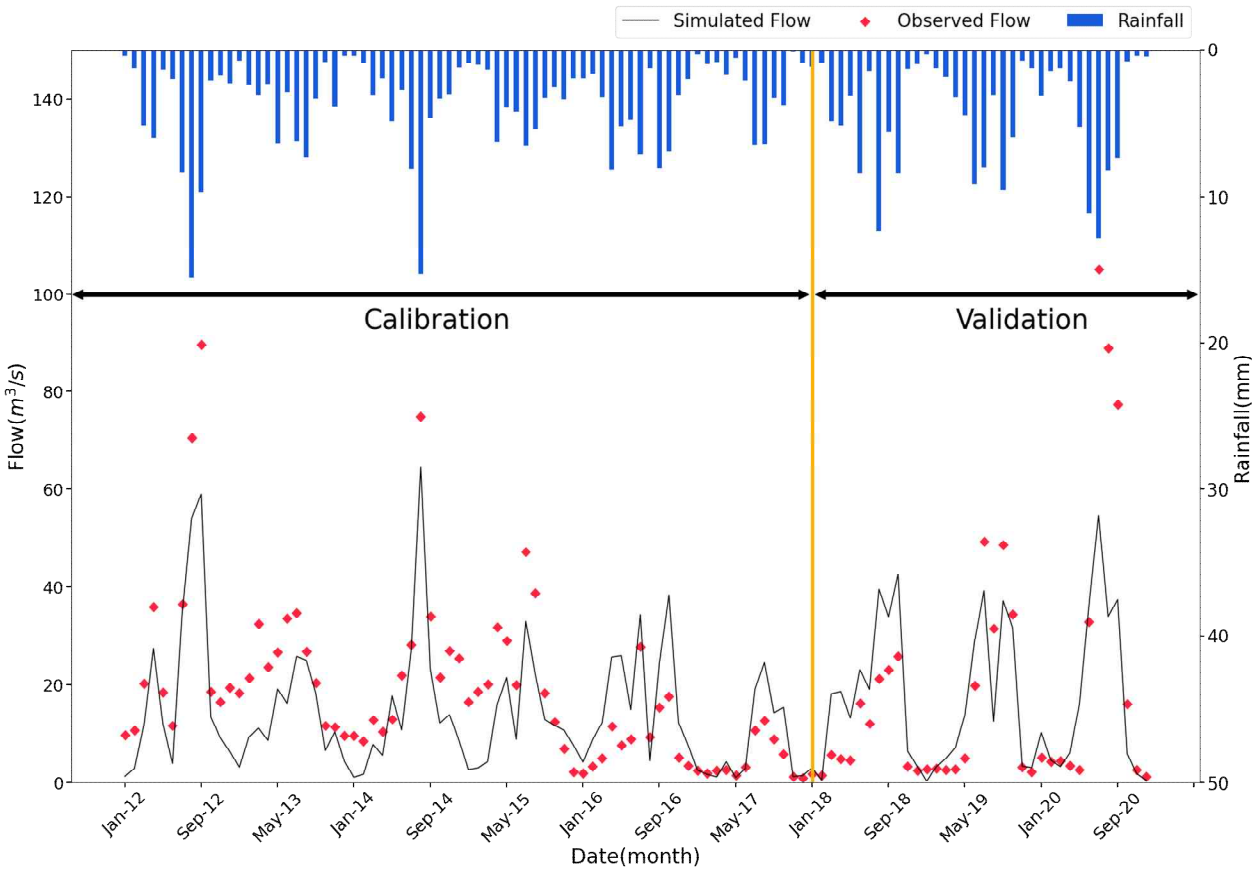
$$RSR = \frac{[\sqrt{\sum_{i=1}^{n}(O_i - P_i)^2}]}{[\sqrt{\sum_{i=1}^{n}(O_i - \overline{O_i})^2}]} \quad (6)$$

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

- 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

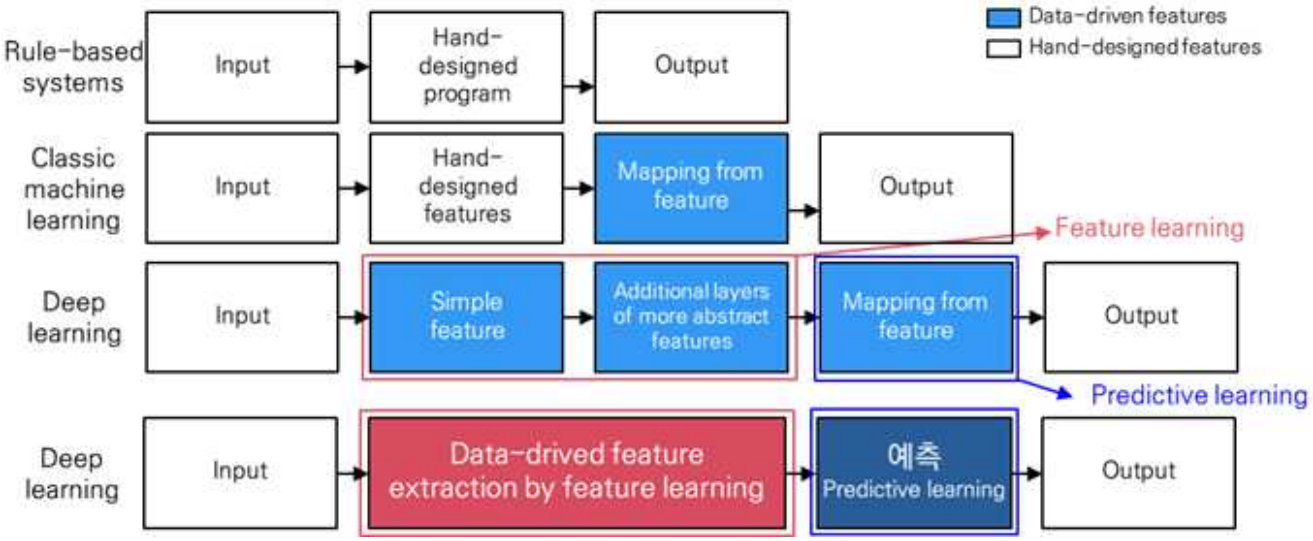


<그림 2-18> SWAT 모형을 활용한 유량 검보정 예시

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.

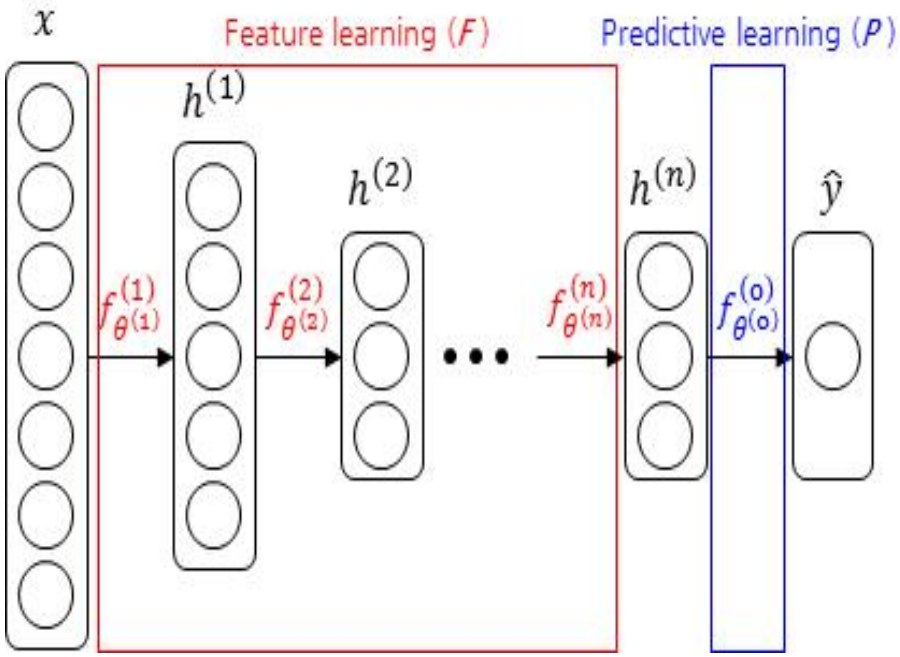


<그림 2-19> 딥러닝 모델의 특징, Goodfellow, 2016

✓ **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